INSTITUT FÜR DEUTSCHE SPRACHE

OPAL

Online publizierte Arbeiten zur Linguistik

ISSN 1860-9422

5/2014

Im Auftrag des Instituts für Deutsche Sprache
herausgegeben von Hardarik Blühdorn, Mechthild Elstermann und Annette Klosa

Heike Stadler

Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora

Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
opal@ids-mannheim.de

Mitglied der Leibniz-Gemeinschaft

Heike Stadler

Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora

Inhalt

Zusammenfassung

Die Basislemmaliste (BLL) der neuhochdeutschen (nhd.) Standardsprache ist eine korpusbasierte, frequenzsortierte Lemmaliste mit mehr als 325.000 Einträgen. Jedes Lemma wird ergänzt durch Wortarten- und Häufigkeitsangaben. Die im Folgenden vorgestellte Version 1.0 der BLL wurde aus DEREKO, dem Deutschen Referenzkorpus des Instituts für Deutsche Sprache, mit 5 Milliarden Wortformen erstellt. Weitere Sprachressourcen sind linguistische Korpusannotationen, die von linguistischen Annotationswerkzeugen wie Lemmatisierern, Part-of-Speech-Taggern oder Parsern stammen. Für die Erstellung der BLL ist das Lemma und das Part-of-Speech-Tag relevant. Die Distanz zwischen lexikografischen Konventionen und maschineller Realität in Form von automatisch vergebenen Lemma-Annotationen erfordert einen Abgleich der aus den Korpusannotationen automatisch generierten Lemmalisten mit der digital verfügbaren Lemmastrecke eines Wörterbuches. Zum einen, um die Vollständigkeit der Einträge frequenter Wörter und das Vorkommen seltener Simplizia in der BLL zu gewährleisten, zum anderen, um die Lemmaform und die Lemmagranularität an die Erwartungen anzupassen, die ein menschlicher Benutzer an ein lexikalisches Verzeichnis der nhd. Standardsprache stellt.

1. Einleitung

Die Basislemmaliste (BLL) der neuhochdeutschen (nhd.) Standardsprache wurde am Institut für Deutsche Sprache (IDS) Mannheim innerhalb des interdisziplinären Projektes *Wechselwirkungen zwischen linguistischen und bioinformatischen Methoden, Verfahren und Algorithmen: Varianz in Sprache und Genomen* konzipiert und erstellt.[1] Die Bestimmung der Analogien und Differenzen von Varianz in Sprache und Genomen basiert auf der Identifikation, Abbildung und Analyse homogener und heterogener Strukturen und Mechanismen in beiden Systemen.[2] Im Mittelpunkt der Abbildung und Analyse sprachlicher Varianz standen Wörter und ihr Wandel in Raum und Zeit. Lexikalische Varianz wird im Projektkontext durch die BLL als Referenz für den allgemeinen öffentlichen Gebrauch der nhd. Standardsprache präsentiert sowie durch Lemmastrecken verschiedener Wörterbücher aus dem Trierer Wörterbuchnetz[3], die die Lexik dialektaler oder diachroner Varietäten des Deutschen enthalten.[4]

Im Wechselwirkungs-Projekt fungiert die BLL als gemeinsames Drittes für die Lemmata der Varietäten-Wörterbücher, die in einer Metalemmaliste[5] mit der BLL verknüpft sind. Die Verbindung zwischen Varietät und nhd. Standardform wird über verschiedene Arten der automatischen Alignierung generiert. Die Kodierung der sprachlichen Vernetzung in der Metalemmaliste erfolgt in XML nach den Richtlinien der *Text Encoding Initiative* (TEI).[6] Die aktuelle Version 1.0 der BLL mit mehr als 325.000 Lemmata mit POS- und Frequenzangaben wird unter DEREWO[7] – dem IDS-Portal für korpusbasierte Grund- und Wortformenlisten – frei zur Verfügung gestellt.

---

[1] Das Verbundprojekt *Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Varianz in Sprache und Genomen* wurde im Rahmen des Förderschwerpunktes *Wechselwirkungen zwischen Natur- und Geisteswissenschaften* vom Bundesministerium für Bildung und Forschung unter der Kennziffer 01UA0815 gefördert.

[2] Die Webseite des Verbundprojektes (www.sprache-und-genome.de/) informiert über Projektpartner, Projektziele und Projektverlauf. Die Schlussberichte sind über die Technische Informationsbibliothek Hannover erhältlich (www.tib.uni-hannover.de/).

[3] Trierer Wörterbuchnetz: http://woerterbuchnetz.de/.

[4] Im Wechselwirkungs-Projekt spielten bei der Abbildung und Analyse sprachlicher und genomischer Varianz auch Morphem- bzw. Domänennetzwerke eine Rolle (Keller/Schultz 2013; Seipel/Borek 2012) .

[5] Metalemmaliste: www.metalemmaliste.de.

[6] Vgl. Seipel/Wegstein (2011). TEI-P5 Richtlinien: www.tei-c.org/Guidelines/P5/.

[7] Perkuhn, Rainer et al. (2012): DEREWO: Korpusbasierte Grundformenliste. www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip.

In ihrer Funktion als Tertium Comparationis für die Lemmata der Varietäten-Wörterbücher soll die BLL zum einen möglichst viele Lemmata der nhd. Standardsprache enthalten, zum anderen sollen Stichwortkonzeption und Lemmaformen den lexikografischen Konventionen entsprechen. Die BLL ist korpusbasiert entstanden, damit sie den tatsächlichen Sprachgebrauch repräsentiert, ihre Grundlage bildet daher fortlaufender Text mit Wortformen. Den Wortformen werden mit Hilfe von linguistischen Annotationswerkzeugen (NLP-Tools) automatisch linguistische Informationen hinzugefügt, beispielsweise das Lemma, die Wortart, Angaben zur Flexion oder Morphologie sowie zu syntaktischen Zusammenhängen.

Für die Erstellung der BLL ist das der Wortform im Korpus zugeordnete Lemma und das Part-of-Speech-Tag (POS-Tag) relevant. Die Lemmalisten der einzelnen NLP-Tools, die aus den linguistisch annotierten Korpora automatisch extrahiert werden, unterscheiden sich untereinander beträchtlich und korrelieren nur bedingt mit den in Wörterbüchern oder Grundwortschätzen verzeichneten Einträgen. Die verschiedenen linguistischen Theorien und sprachlichen Interpretationen, die in die Konzeption und Programmierung der NLP-Tools eingeflossen sind, äußern sich auch in einer abweichenden Zuweisung von POS-Tags an die Wortformen.

Der folgende Abschnitt 2 gibt zunächst eine kurze Definition des Lemmabegriffs, einen Überblick über korpusbasiert entstandene Grundwortschätze, Frequenzwörterbücher sowie Wort(grund)formenlisten und beschreibt Format und Inhalt der BLL. In Abschnitt 3 werden die Sprachressourcen näher vorgestellt, die bei der automatischen Generierung der BLL Verwendung finden. Abschnitt 4 behandelt neben den Inkonsistenzen, die zwischen den Lemmalisten der einzelnen NLP-Tools bestehen, auch deren übereinstimmenden Abweichungen gegenüber dem Lemmabegriff in der Lexikografie sowie die Fehlermöglichkeiten, die sich aus einer automatischen korpusbasierten Darstellung des deutschen Wortschatzes ergeben. Die einzelnen Schritte der Verarbeitung von den Rohlemmalisten der einzelnen NLP-Tools bis zur fertigen BLL werden im Anhang illustriert.

2. Lemma und Frequenz

2.1 Der Lemmabegriff

Als ‘Lemma’ wird in der Lexikografie das Stichwort in einem Nachschlagewerk bezeichnet. Die Zitierform im Wörterbuch ist üblicherweise identisch mit der Wortgrundform, auf die die flektierten Formen eines Wortes zurückgeführt werden. Für Nomen entspricht die Grundform im Deutschen dem Nominativ Singular, für Verben dem Infinitiv und für Adjektive dem Positiv. Die Reduktion der Flexionsformen eines Wortes auf die Wortgrundform wird ‘Lemmatisierung’ genannt. Die Lemmata der Stichwortliste im Wörterbuch werden durch ‘Lemmaselektion’ ermittelt. Lemmatisierungskonzepte aus lexikografischer Sicht werden in folgenden Artikeln eingehend erläutert: Gallmann (1991), Knowles/Zuraidah (2004), Rosengren (1969), Schnörch (2005), Wiegand (1983).

Die automatische Zuordnung einer Grundform zu einer Wortform in der maschinellen Sprachverarbeitung wird bestimmt von den Lexika, Morphologiekomponenten und Algorithmen, mit denen die linguistischen Annotationswerkzeuge intern arbeiten, sowie den Kontextinformationen, die sie bei der Lemmatisierung integrieren. In Abhängigkeit zum Auftreten der Wortform im Lexikon oder der Generierbarkeit der Wortform durch die morphologischen Komponenten wird ein Lemma an die Wortform vergeben, oder nicht. Im

Falle der Nichterkennung wird der Wortform kein Lemma zugewiesen, eine Option, die in den lexikografischen Publikationen zur Lemmatisierung nicht auftritt. Die genauen Lemmatisierungskonzepte der NLP-Tools werden in den zugehörigen Dokumentationen und Publikationen nicht erläutert und die Lexika und morphologischen Komponenten sind in verschlüsselter Form im Quellcode enthalten. Die Fehleranalyse der automatischen Lemmatisierung kann nur anhand der vergebenen Lemmata erfolgen.

Im Weiteren werden als 'Lemmata' sowohl die Stichwörter in einem Wörterbuch oder in einer Wortgrundformenliste bezeichnet, als auch die Grundformen, die den Wortformen im Korpus während der linguistischen Annotation durch ein NLP-Tool zugeordnet werden (Stichwortkandidaten). Die maschinelle Rückführung von Wortformen auf deren Lemmaformen durch ein linguistisches Annotationswerkzeug unterscheidet sich mitunter erheblich von den lexikografischen Konventionen. Die maschinelle Umsetzung der lexikografischen Vorgaben zur Lemmatisierung zeigt ein mannigfaltiges Spektrum an Interpretations- und Fehlermöglichkeiten (vgl. Abschnitt 4). Zur Erstellung einer unter lexikografischen Gesichtspunkten validen korpusbasierten Wortgrundformenliste ist ein Abgleich der aus den linguistisch annotierten Korpora extrahierten Lemmalisten mit der digital verfügbaren Lemmastrecke eines Wörterbuches unumgänglich.

2.2 Häufigkeitswörterbücher, Grundwortschätze und Wort(grund)formenlisten

Die Anzahl der tatsächlichen Realisierungen von Wörtern in der geschriebenen oder gesprochenen Sprache bildet die Grundlage von unterschiedlichen lexikografischen Nachschlagewerken. Frequenzwörterbücher, Wort(grund)formenlisten und häufig auch Grundwortschätze sind korpusbasiert entstanden. Die Häufigkeit der Wortformen dient als Auswahlkriterium für die Aufnahme der Stichworte in die Stichwortliste und ist maßgeblich für die Repräsentationsform. Bei der Erstellung von Grundwortschätzen spielt die didaktische Perspektive eine große Rolle,[8] die häufigsten Wörter werden unter dem Gesichtspunkt ihrer Relevanz im Zweitspracherwerb händisch vervollständigt oder aussortiert. Bei Frequenzwörterbüchern und Wort(grund)formenlisten hingegen entscheidet primär die Anzahl der Vorkommen von Wortformen, bzw. deren Akkumulation zu Wortgrundformen, über die Aufnahme in die Stichwortliste.

Das Interesse an Sprachkorpora und an Frequenzen von Wortformen wurde in der Linguistik durch die Möglichkeit der maschinellen Verarbeitung erheblich gesteigert. Zwei frühe Häufigkeitswörterbücher verdeutlichen, welcher Stellenwert rechnerisch ermittelten Informationen in den Anfangszeiten der angewandten Informatik beigemessen wurde. Rosengren (1972-1977) untersuchte die deutsche Zeitungssprache. Verschlagwortet wurden im 1. Band des Frequenzwörterbuchs die einzelnen Wortformen, im 2. Band die zugehörigen Wortgrundformen. Neben absoluten und relativen Frequenzen werden zu jedem Eintrag auch Ergebnisse des Chi-Quadrat-Tests und eines Streuungsmaßes, die korrigierten Frequenzen (ein zusammenfassendes Maß für Streuung und Frequenz) sowie die Verteilung der Wörter auf die Autoren und Artikel gezeigt. Eine detaillierte Auflistung der Frequenzen nach Themengebieten der Zeitungen rundet die lexikografische Darstellung ab. Die tabellarische Veranschaulichung erfolgt für jede der beiden das Korpus konstituierenden Zeitungen separat. Das Häufigkeitswörterbuch von Ruoff (1981) zur gesprochenen Sprache verzeichnet nur Wort-

[8] Strategien zur Ermittlung, Analyse und lexikografischen Aufbereitung des zentralen Deutschen Wortschatzes sowie die Konzeption ausgewählter Grundwortschätze und Häufigkeitswörterbücher werden in Schnörch (2002) beschrieben.

grundformen, die, nach ihrer Wortartenzugehörigkeit geordnet, alphabetisch, rückläufig-alphabetisch oder numerisch sortiert sind. Zu jeder Wortgrundform werden die absolute und relative Frequenz angegeben.

Zum Entstehungszeitpunkt der Häufigkeitswörterbücher waren die Korpusbasiertheit, die systematische Berechnung statistischer Angaben für eine große Anzahl von Wortformen und die Sortierung der Einträge nach unterschiedlichen Kriterien ein Novum. Rosengren (1972-1977) benutzte für die Auswertung des auf Lochstreifen vorliegenden Zeitungskorpus eine ganze Rechenanlage. Ruoff (1981) stützte sein Wörterbuch auf Tonbandaufnahmen gesprochener Sprache, die sprachlichen Belege wurden fortlaufend in EDV-lesbarer Schrift (OCR-Kugelkopf) archiviert. Die eigentlichen EDV-Arbeiten, die die Berechnung der Frequenzen und die Sortierung der Belege beinhalten, wurden von Experten in einem Rechenzentrum ausgeführt. Die Lemmatisierung und die Wortartenbestimmung erfolgten bei Rosengren und bei Ruoff per Hand. Heute werden entsprechende lexikografische Verzeichnisse eher als Listen bezeichnet denn als Wörterbücher, ähnliche Angaben werden beispielsweise in den Frequenzlisten von Leech/Rayson/Wilson (2001) zum *British National Corpus* gemacht.

In dem 2006 erschienenen Frequenzwörterbuch des deutschen Kernwortschatzes für den Fremdspracherwerb von Randall/Tschirner bildet eine automatische Lemmatisierung und Wortartenbestimmung erstmalig die Grundlage für die Erstellung eines deutschen Grundwortschatzes, neben *TreeTagger* wird das *Wordsmith-Tool* Version 3 angewandt. Trotz der sich daraus ergebenden Möglichkeit der umfangreichen Nutzung der linguistischen Annotationen wird die Disambiguierung von Wortformen und die Lemmatisierung größtenteils per Hand erstellt (Tschirner 2005: 139). Der Korpusumfang wird auf 4,2 Millionen Wortformen festgelegt, damit es „noch per Hand bearbeitet werden konnte und dessen Wortlisten lemmatisiert werden konnten“ (ebd.: 137). Frequenzangaben werden im Kernwortschatz nicht gemacht, die Einträge werden in absteigender numerischer Reihenfolge präsentiert und durch das POS-Tag, eine englische Übersetzung und einen Beispielsatz ergänzt. Die Kriterien für die Auswahl und Vervollständigung der nach dem Frequenzkriterium gewonnenen Stichwörter werden im Vorwort erläutert.

Die Anzahl der von Randall/Tschirner aufgeführten Stichwörter beträgt etwas über 4.000. Aufgenommen werden Grundformen, die mindestens 16-mal im Korpus vorkommen. Das Zeitungskorpus, das Rosengren (1972-1977) als Grundlage diente, bestand aus 3 Millionen Wortformen. Die Liste B der Wörterbücher wurde nach einem rein frequenzbasierten Aufnahmekriterium erstellt, die untere Grenze für die Aufnahme in die Liste liegt bei fünf Vorkommen im Korpus. Liste B des Frequenzwörterbuchs enthält für die *Welt* 31.703 Wortformen und 25.018 Wortgrundformen bzw. für die *Süddeutsche Zeitung* 9.805 Wortformen und 8.426 Wortgrundformen. Wortformen, die dem Flexionsparadigma des häufigsten Lemmas *der* angehören, stehen 279.120-mal im Korpus. Das Korpus von Ruoff (1981) bestand aus 500.000 Wörtern transkribierter gesprochener Sprache, aufgenommen in das Wörterbuch wurden auch Hapax legomena. Das Häufigkeitswörterbuch verzeichnet 15.676 Wortgrundformen, die häufigste Grundform *der*/*dieser* ist 37.536-mal im Korpus belegt.

Die 2009 erschienene korpusbasierte Wortgrundformenliste in der Reihe DeReWo[9] des IDS Mannheim verzeichnet 40.000 Einträge ohne Wortartenangaben. In einem ersten Schritt wurde eine 800.000 Einträge umfassende Wortformenliste aus dem Deutschen Referenzkorpus

[9] DeReWo: www.ids-mannheim.de/kl/projekte/methoden/derewo.html.

DEREKO[10] erstellt, deren Einträge anschließend mit dem Lemmatisierer *glemm* automatisch auf Lemmata abgebildet wurden. Die lexikografischen Kriterien für die händische Nachbearbeitung der mit *glemm* generierten Lemmaliste zur Überführung derselben in die Wortgrundformliste des Deutschen sind in den „Allgemeinen Anmerkungen" zur DEREWO-Reihe ausführlich dokumentiert.

2.3 Die Basislemmaliste

Mit der vollständigen linguistischen Annotation von DEREKO durch mehrere NLP-Tools und der Verfügbarkeit der Stichwortliste von *elexiko*,[11] dem Online-Wörterbuch der Deutschen Gegenwartssprache, besteht die Möglichkeit, eine frequenzsortierte Basislemmaliste mit Wortartenabgaben der nhd. Standardsprache weitgehend maschinell zu generieren. Die BLL der nhd. Standardsprache Version 1.0 (2012) verzeichnet mehr als 325.000 Einträge. Die BLL stellt keine einfache Abbildung der maschinellen Lemmatisierung der Wortformen im Korpus bereit, sie soll häufige und lexikografisch relevante Wörter des Deutschen enthalten. Die lexikografische Validität der automatisch generierten Einträge wird über den Abgleich mit der Stichwortliste von *elexiko* garantiert, ungeeignete Lemmata werden aussortiert (Tokens ohne lexikalische Bedeutung, Orthographiefehler, Lemmatisierungsfehler, fremdsprachliches Material, flektierte Wortformen etc.). Darüber hinaus sichert der Vergleich mit der Stichwortliste eines Wörterbuchs die Vollständigkeit der BLL. Der Bezug zum aktuellen Sprachgebrauch wird durch die händische Verifizierung zusätzlicher frequenter Lemmata aus dem Korpus gewährleistet.

Lemmata, die häufig im Korpus vorkommen, jedoch nicht in der Stichwortliste von *elexiko* enthalten sind, werden in diese nach einer manuellen Überprüfung sukzessive eingegliedert. Auf diesem Wege entsteht die Wechselwirkungs-Stichwortliste (WW-Stichwortliste), die neben den Stichwörtern von *elexiko* ca. 25.000 häufige Lemmata aus dem Korpus enthält. Die im Korpus ermittelten zahlenmäßig relevanten Lemmata, die nicht in *elexiko* zu finden sind, stehen durch die permanente Erweiterung der Stichwortliste in der BLL. Die Kriterien für die Aufnahme von Lemmata in die WW-Stichwortliste korrespondieren im Wesentlichen mit den lexikografischen Richtlinien, die der Erstellung der Stichwortliste von *elexiko* zu Grunde liegen.[12]

Eigennamen, Abkürzungen und Buchstaben sind in der BLL nicht enthalten. In Wörterbüchern werden Eigennamen und Abkürzungen entweder in der fortlaufenden alphabetischen Stichwortliste oder in separaten Listen verzeichnet. Durch den großen Anteil von Zeitungstexten in DEREKO (vgl. Abschnitt 3.1) sind Eigennamen und Abkürzungen in den Lemmalisten stark überpräsentiert im Vergleich zu ihrem Anteil an der Lexik in Wörterbüchern. Sie werden mit Hilfe der POS-Tags und regulärer Ausdrücke aus den Lemmalisten der einzelnen NLP-Tools ausgefiltert. Eine zusätzliche Eigennamenliste mit 68.000 Einträgen wird als Ergänzung zur BLL zur Verfügung gestellt. Die automatische Lemmatisierung und das POS-Tagging von Abkürzungen fällt zu heterogen aus, um eine automatisch generierte Abkürzungsliste zu erzeugen. Einzelne alphabetische Zeichen werden mit den unterschiedlichsten POS-Tags getaggt, deren inkonsistente Vergabe verlässliche Aussagen über den tatsächlichen Gebrauch einzelner Buchstaben im Text nicht zulässt. Abkürzungen (Akronyme

[10] DEREKO: www.ids-mannheim.de/kl/projekte/korpora/, Kupietz/Belica/Keibel/Witt (2010); Kupietz/Keibel (2009), vgl. Abschnitt 3.1.

[11] *elexiko*: www.owid.de/wb/elexiko/start.html, Haß (2005); Klosa (2011) (vgl. Abschnitt 3.3).

[12] In den beiden Sammelbänden zu *elexiko* von Haß (2005) und Klosa (2011) sind für die Konzeption der Stichwortliste besonders die Artikel von Schnörch (2005), Erb (2005), Klosa (2005, 2011) und Klosa/Schoolaert (2011) interessant.

der Schriftsprache) und Buchstaben werden aus den Lemmalisten lediglich ausgegliedert. Kurzwörter (*Alu*, *Auto*, *Lkw*, *GPS*, *Kfz*) bleiben erhalten, sie werden von den NLP-Tools mit einem POS-Tag ausgezeichnet, das der Wortart ihrer Langformen entspricht, sie stehen als Lemmata in der BLL.

Die BLL der nhd. Standardsprache ist absteigend nach der Häufigkeit der Lemmata sortiert und enthält zu jedem Lemma ein POS-Tag sowie den Rang und die Häufigkeitsklasse der Lemma-POS-Tag Kombination (vgl. Abschnitt 2.4). Innerhalb derselben absoluten Frequenz ist die BLL alphabetisch sortiert.

Rang	Lemma	POS	Häufigkeitsklasse
1	der,die,das	ART	0
2	in	PREP	2
3	und	CONJ	2
4	sein	V	3
5	ein(e)	ART	3
6	werden	V	3
7	von	PREP	3
8	mit	PREP	3
9	der,die,das	PRON	3
10	haben	V	3
11	im	PREP	4
...			
326933	zerscherben	V	29
326934	Zetetikerin	NN	29
326935	Zökum	NN	29
326936	zollhoch	ADJ	29
326937	zwerch	ADJ	29
326938	Zwergewerfen	NN	29
326939	zwiegeschlechtig	ADJ	29
326940	Zwiegeschlechtigkeit	NN	29
326941	Zwiegeschlechtlichkeit	NN	29
326942	Zwölfflach	NN	29
326943	zyanenblau	ADJ	29
326944	Zyansäure	NN	29
326945	Zyklogramm	NN	29
326946	Zyrtolith	NN	29

Tab. 1: BLL 1.0

Lemma und POS-Tag bilden eine Einheit, damit ambige Lemmata eindeutig zu bestimmen sind. Unter einem ambigen Lemma wird ein Lemma mit mehreren POS-Tags verstanden. Die Auflösung von Homographen erfolgt ausschließlich über das POS-Tag. Sowohl Homographen, die ihre Verschiedenheit nur durch die Aussprache markieren (*Tenor*, *Tenor*), als auch Homographen, die eine abweichende morphologische Zerlegung aufweisen (*Stau-becken*, *Staub-ecken*), werden unter einem Lemma zusammengefasst. Ebenso erhalten Polyseme nur einen Eintrag in der BLL.

4	sein	V	3	
23	sein	PRON	4	
685	modern	ADJ	10	
20619	modern	V	16	
6537	Hirsch	NN	13	
2127	Hirsch	NE	14	(Eigennamenliste)

Tab. 2: Ambige Lemmata aus der BLL 1.0 und der Eigennamenliste

2.4 Häufigkeitsklassen

Die Verwendung von Häufigkeitsklassen orientiert sich an der in der DEREWO-Reihe üblichen Maßangabe für Frequenzen. Häufigkeitsklassen sollen von den absoluten Frequenzen abstrahieren, die je nach Korpusgrundlage und NLP-Tool stark variieren können. Die Häufigkeitsklasse *N* eines Lemmas berechnet sich aus der absoluten Frequenz *F* des untersuchten Lemmas im Verhältnis zur absoluten Frequenz *F* des häufigsten Lemmas:

$$N = \mathrm{HK}(\textit{untersuchtes Lemma}) := \left[\log_2 \left(\frac{F(\textit{häufigstes Lemma})}{F(\textit{untersuchtes Lemma})} \right) + 0{,}5 \right]$$

Formel 1: Berechnung der Häufigkeitsklasse

Die Anzahl der Häufigkeitsklassen richtet sich nach dem Vorkommen des häufigsten Lemmas im Korpus. Bei einem großen Korpus ist die ermittelte Frequenz des häufigsten Lemmas sehr hoch, und damit auch die Anzahl der Häufigkeitsklassen größer, als bei einem kleinen Korpus. Die Differenz in der Anzahl der Häufigkeitsklassen ist jedoch sehr viel geringer als die Differenz der Frequenzen in absoluten Zahlen.

Die BLL 1.0 basiert auf dem Deutschen Referenzkorpus DEREKO, das mehr als 5 Milliarden Wortformen enthält (vgl. Abschnitt 3.1). Das häufigste Lemma der BLL 1.0 *der,die,das* mit dem POS-Tag ART hat in diesem Korpus die Frequenz 490.092.568. Für ein Hapax legomenon wird in der BLL 1.0 die Häufigkeitsklasse 29 berechnet. In einem Korpus bestehend aus nur 1 Million Wortformen kommt das häufigste Lemma ca. 100.000-mal vor. In einer zugehörigen Lemmaliste sind 17 Häufigkeitsklassen vertreten:

$$29 = \left[\log_2 \left(\frac{490.092.568}{1} \right) + 0{,}5 \right] \qquad 17 = \left[\log_2 \left(\frac{100.000}{1} \right) + 0{,}5 \right]$$

Abb. 1: Berechnung der Häufigkeitsklassen eines Hapax legomenon in einem Korpus mit 5 Milliarden Tokens (häufigstes Lemma 490.092.568 Vorkommen) bzw. 1 Million Tokens (häufigstes Lemma ca. 100.000 Vorkommen)

Die Frequenzen der Lemmata einer bestimmten Häufigkeitsklasse *N* liegen innerhalb eines bestimmten Intervalls:

$$\left[\left(\frac{F(\textit{häufigstes Lemma})}{2^{(N-0,5)}} \right) \right] \geq F(\textit{Lemmata} \in N) > \left[\left(\frac{F(\textit{häufigstes Lemma})}{2^{(N+0,5)}} \right) \right]$$

Formel 2: Berechnung des Frequenzbereichs einer Häufigkeitsklasse *N*

Das häufigste Lemma einer frequenzsortierten Liste ist immer in der Häufigkeitsklasse 0 (Beispiel: $[(490.092.568 / 2^0) + 0{,}5] = 490.092.568$). Die Frequenz der Lemmata in der höchsten Häufigkeitsklasse beträgt immer 1 (Beispiel: $[(490.092.568 / 2^{29}) + 0{,}5] = 1$). In den niedrigen Häufigkeitsklassen liegen sehr viele mögliche Frequenzen (Beispiel: $693.095.556 \geq F(\textit{Lemmata} \in \textit{HK}\ 0) > 346.547.778$), während in den hohen Häufigkeitsklassen die Anzahl der möglichen Frequenzen wesentlich geringer ist (Beispiel: $1 \geq F(\textit{Lemmata} \in \textit{HK}\ 29) > 0$). Die Anzahl der Lemmata, die den Häufigkeitsklassen angehören, verhält sich zu dieser Tatsache umgekehrt proportional. In den kleinen Häufigkeitsklassen sind nur sehr wenige Lemmata verzeichnet, die hohen Häufigkeitsklassen umfassen den Großteil des Wortschatzes (vgl. Abb. 2, Tab. 1).

In den Häufigkeitsklassen spiegelt sich die von Zipf (1932, 1935) formulierte Verteilung des Wortschatzes wider, wonach sich die Frequenz *f* eines Wortes umgekehrt proportional zu dessen Rang *r* innerhalb einer absteigend frequenzsortierten Wortliste verhält. Das Produkt aus Frequenz und Rang der Wörter ist annähernd konstant.

$$f \propto \frac{1}{r} \qquad f \cdot r = k$$

Formel 3: Zipfs Gesetz

Häufigkeitsklassen werden auch in den monolingualen, korpusbasierten Wörterbüchern des Wortschatz-Portals der Uni Leipzig neben der absoluten Frequenz als numerische Information zu einem Lemma gegeben.[13] Unter den „Fragen zu den Suchergebnissen“ des Wortschatz-Portals wird der Zusammenhang zwischen Zipf'schem Gesetz und Häufigkeitsklassen anschaulich erläutert.[14] Eine ausführliche Darstellung von Häufigkeitsklassen und weiteren in der Korpuslinguistik gebräuchlichen Häufigkeitsmaßen geben Perkuhn/Keibel/Kupietz (2012). Wortfrequenz- und Zipf-Verteilung sind Gegenstand zahlreicher Monografien und Sammelbände der Quantitativen Linguistik (z.B. Guiter/Arapov 1982; Baayen 2001; Popescu 2009).

Zur Berechnung der Häufigkeitsklassen der im Wechselwirkungs-Projekt als separate Liste erstellten Eigennamenliste wird neben dem untersuchten Lemma die Frequenz des häufigsten Lemmas der BLL verwendet. Die Eigennamenliste ist eine Teilliste der BLL bestehend aus Lemmata mit dem POS-Tag *NE* (‘Named Entitiy’) (vgl. Tab. 26). Beide Listen werden aus demselben Korpus gebildet.

Wenn zwei zu vergleichende Lemmalisten nicht die gleiche Korpusgrundlage haben, werden die Häufigkeiten beider Listen zunächst zu einer bestimmten Zahl ins Verhältnis gesetzt und die Häufigkeitsklassen danach mit den angepassten Frequenzen errechnet, um äquivalente Ergebnisse zu erzielen. Die Normalisierung von Frequenzen in Lemmalisten, die aus Korpora unterschiedlicher Größe stammen, wenden z.B. Leech/Rayson/Wilson (2001) an. Sie vergleichen Lemmalisten, die auf der Grundlage von geschriebener und gesprochener Sprache des *British National Corpus* entstanden, und berechnen gerundete Frequenzen (*rounded frequency*), indem sie die absoluten Frequenzen auf ihre proportionale Häufigkeit in einer Million Tokens glätten.

2.5 Frequenzkriterium und lexikografische Relevanz

Mit mehr als 325.000 Lemmata besitzt die BLL 1.0 deutlich mehr Einträge als bisherige Häufigkeitswörterbücher und Wortgrundformenlisten für das Deutsche. Die automatische Lemmatisierung der Wortformen und das POS-Tagging ermöglichen die Bearbeitung großer Textkorpora. Die automatisch generierten Lemmata und POS-Tags korrelieren jedoch nur bedingt mit dem lexikografischen Standard, die Lemmatisierungskonzepte und Lemmaformen der linguistischen Annotationswerkzeuge unterscheiden sich von den üblichen Konventionen für die Erstellung von Wörterbüchern (vgl. Abschnitt 4). Durch den automatischen Abgleich mit der Stichwortliste von *elexiko* wird die lexikografische Validität der Lemmata sichergestellt, die in die BLL einfließen. Die permanente Erweiterung der Stichwortliste von *elexiko*

[13] Wortschatz-Portal – monolinguale Wörterbücher: http://corpora.informatik.uni-leipzig.de/.
[14] Wortschatz-Portal – Häufigkeitsklassen: http://asvdoku.informatik.uni-leipzig.de/corpora/index.php?id=fragen-zu-den-suchergebnissen.

während der Projektlaufzeit zur WW-Stichwortliste durch ca. 25.000 händisch verifizierte häufige Lemmata aus dem Korpus gewährleistet die Repräsentativität der BLL für die aktuelle deutsche Standardsprache.

Der Abgleich der automatisch generierten Lemmalisten der einzelnen NLP-Tools mit einer Stichwortliste reduziert die Frequenz jedoch auf die Sortiermöglichkeit einer Liste, deren Einträge a priori feststehen. Als Aufnahmekriterium für ein Lemma aus dem Korpus in die BLL dient nicht dessen Frequenz, sondern das Vorkommen des Lemmas in der WW-Stichwortliste. Die automatisch generierten frequenzsortierten Lemmalisten der NLP-Tools sind zur Darstellung des standardsprachlichen deutschen Wortschatzes nicht geeignet. Sie enthalten aus dem Korpus extrahierte Zeichenkombinationen, die keine Wörter bezeichnen, Orthographiefehler und fremdsprachliches Material sowie durch die automatische Lemmatisierung produzierte Fehler und eine Vielzahl an Lemmata, die lediglich Wortformen einer Flexionsreihe anderer Lemmata darstellen, und unter diesem zu subsumieren sind. Die automatisch aus dem Korpus generierten Lemmalisten entsprechen nicht der nhd. Standardsprache, eine komplette händische Nachbearbeitung der rein frequenzbasierten automatisch generierten Lemmalisten wäre daher unumgänglich.

Im Wechselwirkungs-Projekt bildet die BLL ein Referenzwortverzeichnis für Wörterbücher mit regional und zeitlich markierten Varietätenlemmata. Die Abbildung der nhd. Standardsprache in der BLL ist daher das primäre Ziel. Die jahrelange fundierte lexikografische Arbeit an der *elexiko*-Stichwortliste wurde im Wechselwirkungs-Projekt genutzt, um den Fokus der manuellen Nacharbeiten auf die Verifizierung bzw. Abweisung von Lemmakandidaten aus dem Korpus mit einer hohen Frequenz zu legen, die noch nicht in der *elexiko*-Stichwortliste stehen. Die zusätzlich aus dem Korpus gewonnenen Lemmata repräsentieren in besonderem Maße die Lexik von Themen, die den Sprachgebrauch im Korpus in den vergangenen Jahr(zehnt)en prägten.

In der *elexiko*-Stichwortliste sind des Weiteren Angaben zu den orthographischen Varianten eines Lemmas in der aktuellen Standardschreibweise vorhanden. Diese Informationen werden verwendet, um orthographische Varianten im Korpus zu identifizieren und auf die heutige Standardschreibweise abzubilden. In der BLL stehen die automatisch addierten Frequenzen der Orthographievarianten bei dem Lemma, das dem Standard entspricht und das als einziges in der BLL verzeichnet ist.

Aufgrund der Selektion der Lemmata durch den Abgleich mit einer vorgegebenen Stichwortliste entspricht die Verteilung der Lemmata der BLL nach Häufigkeitsklassen (Abb. 3) nicht der üblichen Häufigkeitsverteilung (Abb. 2) nach dem Zipf'schen Gesetz (vgl. Abschnitt 2.4).

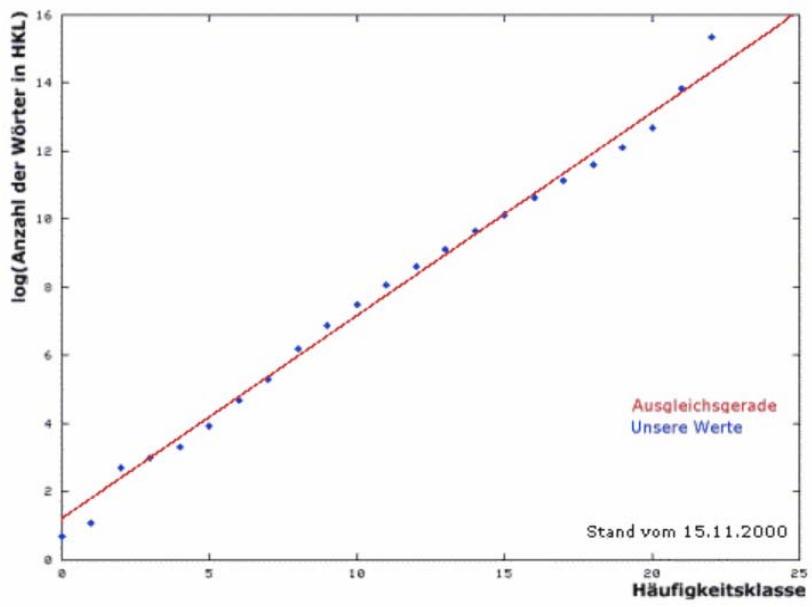

Abb. 2: Verteilung der Wörter nach Häufigkeitsklassen beim Wortschatz-Portal der Uni Leipzig[15]

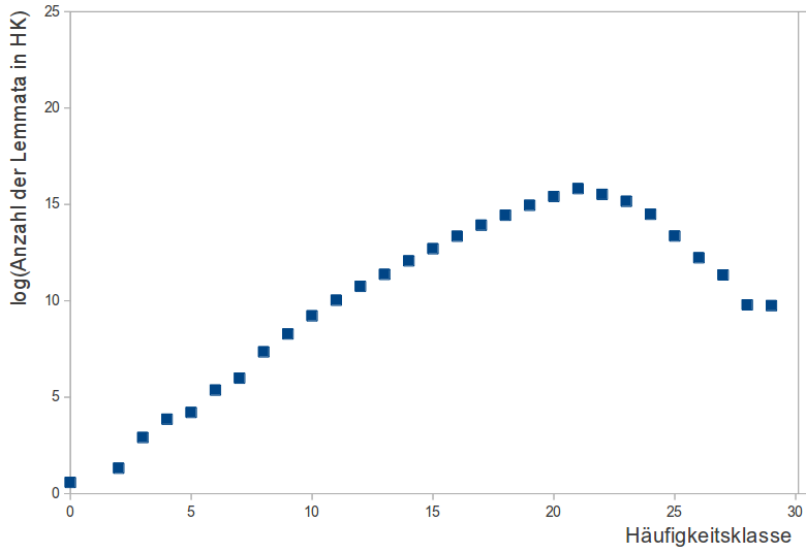

Abb. 3: Verteilung der Lemmata nach Häufigkeitsklassen in der BLL 1.0

Von den zahlreichen Hapax legomena und niedrigfrequenten Lemmata im Korpus stehen nur wenige in der *elexiko*-Stichwortliste. Die pauschale Aufnahme der niedrigfrequenten Lemmata der *elexiko*-Stichwortliste in die BLL wird durch ihre meist gegebene lexikografische Relevanz begründet. Während es sich bei den Lemmata, die aufgrund ihres häufigen Vorkommens zusätzlich zu den Stichwörtern aus *elexiko* in die WW-Stichwortliste aufgenommen werden, fast ausschließlich um transparente Komposita handelt (vgl. Anhang), konstituieren die niedrigfrequenten Lemmata der BLL überwiegend Lexeme, deren Bedeutung nicht aus den einzelnen Bestandteilen erschlossen werden kann. Die Abhängigkeit der Frequenz und Listenposition einzelner Lemmata von den unterschiedlichen Lemmatisierungskonzepten der NLP-Tools wird in Abschnitt 4.1.1 näher erläutert. Kenngrößen wie zeitliche Stabilität, hohe Wortbildungspotenz oder regionale Gleichverteilung eines Lemmas, die ebenfalls auto-

[15] http://asvdoku.informatik.uni-leipzig.de/corpora/index.php?id=fragen-zu-den-suchergebnissen

matisch anhand der Korpusgrundlage ermittelt werden können, spielen in dem im Folgenden vorgestellten Ansatz keine Rolle für die Aufnahme eines Lemmas in die BLL.

3. Sprachressourcen

Bei der Erstellung der BLL werden drei unterschiedliche Arten von Sprachressourcen verwendet:

- Digitale Sprachkorpora
  - DEREKO
- Linguistische Annotationswerkzeuge (NLP-Tools)
  - Lemmatisierer *glemm*
  - POS-Tagger *TreeTagger*, *Machinese Phrase Tagger*
  - Parser *Machinese Syntax*, *Xerox Incremental Parser XIP*
- Digitale Stichwortlisten von Wörterbüchern
  - *elexiko*

DEREKO wurde von mehreren linguistischen Annotationswerkzeugen (NLP-Tools) parallel mit linguistischen Informationen angereichert und die Auszeichnungen in separaten Stand-Off Annotationen im XML-Format gespeichert. Jedes NLP-Tool bestimmt zunächst die Wortformen im Korpus (Tokenisierung) und weist jeder Wortform mindestens ein Lemma und ein POS-Tag zu.

```
<sentence>
    <lexeme id="6" pos="15414" len="3">
        <surface-form>Rot</surface-form>
        <sense id="0">
            <base-form>rot</base-form>
            <part-of-speech conf="0.730213">ADJD</part-of-speech>
        </sense>
        <sense id="1">
            <base-form>Rot</base-form>
            <part-of-speech conf="0.269787">NN</part-of-speech>
        </sense>
    </lexeme>
    …
```

Abb. 4: Linguistische Annotation des *TreeTaggers* für die satzinitiale Wortform *Rot* (St. Galler Tagblatt 2000)

Die in den annotierten Korpora enthaltenen Lemmata und ihre zugehörigen POS-Tags werden mit Perl-Skripten extrahiert und in NLP-Tool-spezifischen Lemmalisten summiert. Über den automatischen Vergleich der NLP-Tool-spezifischen Lemmalisten untereinander und mit der WW-Stichwortliste wird die BLL generiert (vgl. Anhang).

3.1 DEREKO

Als Korpusgrundlage für die Erstellung der BLL dient das gesamte Deutsche Referenzkorpus (DEREKO) des Instituts für Deutsche Sprache; es umfasste zum Entstehungszeitpunkt der

BLL Version 1.0 mehr als 5 Milliarden Wortformen.[16] Die öffentliche Schnittstelle zu DeReKo ist COSMAS-II, das *Corpus Search, Management and Analysis System* des IDS Mannheim.[17] Über COSMAS-II ist eine Auswahl aus den DeReKo-Daten abfragbar, die BLL 1.0 basiert auf dem gesamten in DeReKo enthaltenem Sprachmaterial. Die Korpusauszeichnungen in DeReKo werden unter dem Gesichtspunkt ihres Nutzens für die Lexikographie in Klosa/Kupietz/Lüngen (2012) behandelt und nach inhaltlichen Gesichtspunkten gruppiert: Annotation der hierarchischen Korpus- und Textstruktur, linguistisches Tagging, bibliografische Metadaten, textlinguistische Klassifikation, textspezifische deskriptiv-statische Metadaten und textspezifische Metadaten zu Duplikaten. Für die Erstellung der BLL spielen ausschließlich die linguistischen Annotationen eine Rolle.

DeReKo besteht aufgrund der Urheberrechte überwiegend aus Zeitungstexten aus Deutschland, Österreich und der Schweiz. Daneben sind in geringerem Umfang auch literarische Texte, Gebrauchstexte, fachsprachliche Texte und politische Texte vertreten. Ab 2012 sind die Wikipedia-Artikel und -Diskussionen in das Korpus integriert. Die Schriftsprache in DeReKo ist fast ausschließlich auf eine öffentliche, formelle Gesprächssituation beschränkt, das Korpus enthält keine Tweets, E-Mails, Internetforen etc. Dialektal geprägte Schriftsprache tritt nur sporadisch in Kolumnen und Zitaten der Zeitungskorpora sowie als wörtliche Rede in den literarischen Texten auf. Ausgerichtet auf Kontexte großer Öffentlichkeit repräsentiert DeReKo den Gebrauchsstandard des Neuhochdeutschen seit den 1950er Jahren, mit einem sehr deutlichen Zuwachs der Wortformen ab der Jahrtausendwende, als sich die digitale Archivierung endgültig durchsetzte. Lexikalische Besonderheiten des österreichischen und Schweizer Standarddeutsch (*Paradeiser* Ö, *grillieren* CH), werden in die BLL als selbstständiges Lemma übernommen ohne Kennzeichnung als regionale Standardvariante. Orthographische Varianten (z.B. *ss* für *ß* in der Schweiz) werden unter der bundesdeutschen Standardschreibweise subsumiert.

3.2 Linguistische Annotationswerkzeuge (NLP-Tools)

Bei der Erstellung der BLL werden die Lemmata und POS-Tags verarbeitet, die den Wortformen im Korpus von unterschiedlichen linguistischen Annotationswerkzeugen automatisch zugewiesen wurden. Aus den für jedes NLP-Tool separat gespeicherten linguistischen Korpusauszeichnungen wird im ersten Schritt für jedes NLP-Tool einzeln eine Lemmaliste extrahiert, die die Lemmata mit zugehörigen POS-Tags und die akkumulierten Frequenzen für eine Lemma-POS-Tag-Kombination enthält. Neben dem Lemma und dem POS-Tag können im annotierten Korpus weitere morphologische, morphosyntaktische oder syntaktische Angaben vorhanden sein, die für die Erstellung der BLL jedoch nicht ausgewertet werden.[18] Ein Bezug der Lemmata in den Lemmalisten zu Wortformen in einer bestimmten Position im Korpus ist nicht mehr herzustellen. Zusätzlich gespeichert werden die Wortformen, die das Flexionsparadigma eines Lemmas konstituieren, um Entscheidungen der NLP-Tools bei der Lemmatisierung nachzuvollziehen. Mit der Kombination aus Wortform, Lemma und POS-Tag können alle Korpusstellen gefunden werden, die den Kontext für die Lemmatisierung und Wortartenzuweisung zu einer bestimmten Wortform bieten.

[16] Für die Erstellung der BLL Version 1.0 wurde das DeReKo-Release 2012-I vom 29.2.2012 verwendet. Chronik der DeReKo-Freigaben: www.ids-mannheim.de/kl/projekte/korpora/releases.html.

[17] COSMAS-II: www.ids-mannheim.de/cosmas2/.

[18] Eine Ausnahme bilden Angaben zu Eigennamen und Abkürzungen. Deren Spezifikation wird bei zwei der NLP-Tools unter den *Features* gekennzeichnet. In diesem Fall werden die *Features* ebenfalls ausgewertet und das zugehörige POS-Tag konform zum Tagset der BLL umgewandelt zu ‘NE’ bzw. ‘ABBR’ (vgl. Abschnitt 4.2.1).

Als linguistische Annotationswerkzeuge kamen ein Lemmatisierer, zwei POS-Tagger und zwei Dependenz-Parser zum Einsatz. Der Lemmatisierer *glemm* ist nur IDS-intern verfügbar, eine öffentliche Dokumentation gibt Einblick in Designprinzipien, Implementierung und linguistischen Hintergrund.[19] Von der Firma *Connexor Oy's* werden aus der Reihe *Machinese Products* zwei Softwareprodukte verwendet: *Machinese Phrase Tagger* (*MPT*) und *Machinese Syntax* (*MS*). Die beim Kauf der Software von *Connexor* mitgelieferte Dokumentation ist nicht öffentlich zugänglich, auf dem *Cosmas-II*-Portal des IDS werden die Informationen aus der Dokumentation zum morphologischen und syntaktischen Tagset wiedergegeben.[20] Einen Einblick in die Arbeitsweise der *Machinese* Software von *Connexor* erhält man auf der Firmenwebseite.[21] *TreeTagger* ist ein frei verfügbarer am IMS Stuttgart entwickelter Tagger, der die Wahrscheinlichkeit der POS-Tags mittels Entscheidungsbäumen berechnet. Das Stuttgart-Tübinger Tagset und die Vorgehensweise des *TreeTaggers* sind ausführlich dokumentiert.[22] Für die kommerzielle Software *Xerox Incremental Parser* (*XIP*) gibt es Dokumentationen auf der Firmenwebseite.[23] Dort besteht auch die Möglichkeit, Beispielsätze mit dem regelbasierten Dependenz-Parser zu analysieren.[24] Bei den Parsern liegt der Schwerpunkt der linguistischen Analyse auf der Auszeichnung der syntaktischen Strukturen im Satz, die morphosyntaktische Analyse der Wortformen ist jedoch unabdingbar für die Ermittlung der grammatischen Beziehungen. Mit Lemmata und POS-Tags werden die Wortformen im Korpus von allen fünf NLP-Tools gleichermaßen ausgezeichnet.

Von den NLP-Tools wird der Kontext einer Wortform während der linguistischen Annotation des Korpus in ganz unterschiedlichem Maße evaluiert, um Lemmata und POS-Tags zu vergeben. Der Kontext ist immer relevant, wenn es um die Disambiguierung mehrdeutiger Wortformen geht, eine Wortform kann mitunter den Flexionsreihen mehrerer Lemmata zugeordnet sein und mehr als einer Wortart angehören. Der Lemmatisierer *glemm* erhält eine Wortformenliste als Eingabe und vergibt die POS-Tags anhand von Heuristiken. *TreeTagger* bezieht eine bestimmte Anzahl von Wörtern im linken Kontext in die Analyse mit ein. Die Strategien zu Lemmatisierung und POS-Tagging sind den Dokumentationen der *Connexor*-Werkzeuge und von *XIP* nicht zu entnehmen, für die Annotation von syntaktischen Funktionen und Dependenzstrukturen verwenden die Parser den ganzen Satz. Die Güte der Lemmata und POS-Tags der einzelnen NLP-Tools korreliert jedoch nur bedingt mit der besseren Disambiguierungsmöglichkeit der Wortformen durch die Einbeziehung des Kontextes. Die Morphologiekomponenten der NLP-Tools sind ebenso entscheidend wie die Lexika, mit denen die NLP-Tools intern arbeiten. In den internen Lexika werden Lemmata mit ihren zugehörigen Flexionsreihen und POS-Tags kodiert. Wortformen, die nicht im internen Lexikon stehen und denen auch mit der Morphologiekomponente des NLP-Tools keine gültige morphologische Zerlegung zugeordnet werden kann, werden nicht lemmatisiert.

[19] Belica (1994)
[20] Tagset Morphologie *Machinese*: www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/morph.html.
Tagset Syntax *Machinese*: www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/syntax.html.
[21] *Machinese* Demo: www.connexor.com/nlplib/.
[22] *TreeTagger*: www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html; Schmid (1994, 1995); Schiller/Teufel/Thielen (1995).
[23] *XIP* Dokumentation: http://open.xerox.com/Services/XIPParser/Pages/Using XIP.
[24] *XIP* Demo: http://open.xerox.com/Services/XIPParser/Consume/64.

3.3 *elexiko*

Die Stichwortliste des *Online-Wörterbuchs der Deutschen Gegenwartssprache – elexiko* des IDS Mannheim mit 300.000 Lemmata bildet die Grundlage für die lexikografische Validierung der Lemmata in den NLP-Tool-spezifischen Lemmalisten. *elexiko* ist ab Ende der 1990er Jahre entstanden, als sich mit dem World Wide Web das Potenzial eines Online-Wörterbuchs zeigte. In den beiden Sammelbänden von Haß (2005) und Klosa (2011) wird *elexiko* ausführlich dokumentiert. Ausgangspunkt für die Auswahl der Lemmata in *elexiko* war eine Stichwortkandidatenliste, die korpusbasiert ermittelt wurde. Als Korpus dienten große Teile von DeReKo, die Wortgrundformen wurden den Wortformen vom Lemmatisierer *glemm* automatisch zugewiesen. Die Stichwortkandidatenliste enthält nur Wortgrundformen, die mindestens achtmal im Korpus vorkommen. Sie wurde durch eine redaktionelle lexikologisch-lexikografische Überprüfung und Aufbereitung in die *elexiko*-Stichwortliste überführt.[25] Durch den Vergleich mit Lemmastrecken weiterer Wörterbücher wurde sichergestellt, dass keine wichtigen Stichwörter in *elexiko* fehlen.

elexiko ist als lexikalische Datenbank konzipiert, die Wortartikelstruktur wird mit einer komplexen XML-DTD realisiert (Müller-Spitzer 2005). In *elexiko* gibt es zwei Typen von Wortartikeln, der Typ des Wortartikels eines Stichwortes ist abhängig von dessen Frequenz. Niedrigfrequente Stichwörter werden ausschließlich mit automatisch ermittelten Angaben versehen, hochfrequente Stichwörter werden redaktionell vollständig beschrieben. Die automatisch ermittelten Angaben umfassen Korpusbelege, die Frequenzschicht[26] und Informationen zur Verteilung im Korpus. Einige Wortartikel von niedrigfrequenten Stichwörtern beinhalten lediglich deren normgerechte Schreibung und Worttrennung (*Veloroutenkonzept*, *Vidikon*, *Zackenlitze*, *Zetetiker*). In diesen Fällen ist die automatische Bereitstellung der Angaben aus dem Korpus und deren redaktionelle Überprüfung noch nicht abgeschlossen. Die häufigen Stichwörter werden durch eine lexikografische Bearbeitung nach Lesarten disambiguiert und mit Bedeutungserläuterungen, Kollokationen, Konstruktionen, sinnverwandten Wörtern, Gebrauchsbesonderheiten und grammatischen Informationen veranschaulicht. Für zahlreiche häufige Stichwörter werden außerdem automatisch generierte Wortbildungsprodukte gezeigt.

4. Lexikografische Entscheidungen bei der Erstellung der BLL – NLP-Tool-spezifische Lemmalisten im Vergleich[27]

Im Folgenden wird anhand einiger Beispiele beschrieben, welche Arten von Differenzen zwischen den Lemmata und POS-Tags der NLP-Tool-spezifischen Lemmalisten bestehen und in welchen Gemeinsamkeiten sich diese von lexikografischen Konventionen unterscheiden. Fehler und Probleme, die bei der automatischen Zuweisung von Lemmata und POS-Tags auftreten, werden paradigmatisch verdeutlicht. Die Anpassung der automatisch generierten Lemmata an lexikografische Konventionen wurde bei der Erstellung der BLL durch die Implementierung eines Regelsystems in Perl-Skripten realisiert (vgl. Anhang). Fehlerhafte und unter lexikografischen Gesichtspunkten nicht relevante Lemmata werden entweder auf

[25] Die einzelnen Schritte auf dem Weg von der Stichwortkandidatenliste zur *elexiko*-Stichwortliste werden in Schnörch (2005) detailliert beschrieben.

[26] *elexiko* – Frequenzschichten: www.owid.de/wb/elexiko/glossar/Frequenzschichten.html.

[27] Die Annotationen von *XIP* standen zum Entstehungszeitpunkt der BLL 1.0 (2012) aus lizenzrechtlichen Gründen nicht zur Verfügung. Um die linguistischen Annotationen von *XIP* in den Vergleich der Lemmata und POS-Tags der NLP-Tools mit einzubeziehen, wird das DeReKo-Release 2011-I (Stand 29.3.2011) als Korpusgrundlage für den Vergleich der NLP-Tool-spezifischen Lemmalisten verwendet.

eine Lemmaform in der WW-Stichwortliste abgebildet oder aussortiert. Flexionsreihen werden mitunter in den NLP-Tool-spezifischen Lemmalisten und der WW-Stichwortliste von abweichenden Lemmaformen oder einer unterschiedlichen Anzahl an Lemmata repräsentiert. Für diese Fälle wird ein Lemma angesetzt, welches die feiner gegliederten Analysen subsumiert. Verschiedenartige Lemmaformen für Pronomen und Artikel werden über eine händisch erstellte Liste auf eine festgelegte Lemmaform vereinheitlicht.

4.1 Lemmatisierung

4.1.1 Tokens, Types und Lemmata

Von den NLP-Tools werden bei identischer Korpusgrundlage unterschiedlich viele Tokens erkannt. Die Definition der Zeichen, die zu einem Token zusammengefasst werden, ist von NLP-Tool zu NLP-Tool abweichend.

	Flex (lexikalischer Scanner)[28]	*MPT*	*MS*	*TreeTagger*	*XIP*
Tokens	3.668.905.369	4.369.006.062	4.240.504.860	4.382.940.875	4.398.818.486
Types	12.867.440	22.647.644	24.027.985	24.748.830	24.263.497

Tab. 3: Anzahl der von den einzelnen NLP-Tools im Korpus ermittelten Tokens und Types

Während für die Ermittlung der Anzahl an Tokens und Types das POS-Tag nicht berücksichtigt wird, ist es bei den Lemmalisten ein distinktives Merkmal, das zu einem eigenen Eintrag führt, wenn das Lemma mit mehr als einem POS-Tag im Korpus belegt ist. Lemma und POS-Tag bilden eine Einheit, um die Lesarten der in Bezug auf ihre Wortart ambigen Lemmata zu unterscheiden. Ist im Folgenden von den Lemmata in den Lemmalisten die Rede, ist damit die Lemma-POS-Tag-Kombination gemeint. Die Anzahl der Lemmata in den Lemmalisten ist sehr viel stärker abweichend als die der Tokens und Types.

	glemm	*MPT*	*MS*	*TreeTagger*	*XIP*
Lemmata	6.067.482	18.127.803	20.985.211	702.396	22.186.282
Tokens ohne Lemma	3,7 %	–	5,6 %	4,5 %	3,2 %
Types ohne Lemma	31,8 %	–	55,4 %	91,8 %	37,2 %

Tab. 4: Anzahl der Lemmata pro Lemmaliste, Prozentsatz der Tokens und Types, für die kein Lemma gefunden wurde

[28] Der Lemmatisierer *glemm* erfordert als Eingabe eine Tokenliste, die mit einem beliebigen Tokenisierer aus dem Korpus generiert wird. Als Tokenisierer wurde ein IDS-internes auf Flex basierendes Tool eingesetzt, das die Tokens summiert und als Types mit Frequenzen in einer Liste zusammenfasst. In der generierten Liste sind ausschließlich Types enthalten, die zur Gänze aus alphabetischen Zeichen bestehen, Wortformen, die Zahlen oder Bindestriche enthalten, werden nicht aufgeführt. Auf der einen Seite sind daher sehr viele Types in der Frequenzliste nicht enthalten, die im Deutschen als Wort nicht existieren (*www.motortv.de, #Mathematik*), auf der anderen Seite fehlen allerdings Types, die üblicherweise als Bestandteile des deutschen Wortschatzes gelten (*80-jährig*, *H-Milch*). Aufgrund der lexikalischen Auswahlkriterien des auf Flex basierenden lexikalischen Scanners ist die Anzahl der Types und Tokens nicht mit den Listen der weiteren NLP-Tools vergleichbar. Da *glemm* Wortformen, die Ziffern oder andere nicht alphabetische Zeichen enthalten, nicht analysiert, gehen durch die Verwendung der mit dem lexikalischen Scanner ermittelten Tokenliste keine lexikalischen Informationen verloren.

Die große Differenz zwischen der Anzahl der Lemmata in den Lemmalisten liegt hauptsächlich an der Behandlung von Tokens, die durch die internen Lexika oder Morphologiekomponenten der NLP-Tools nicht als Wortformen erkannt werden. *MPT*, *MS* und *XIP* weisen diesen Tokens das Token selbst als Lemma zu: *MS* zeichnet die betroffenen Lemmata mit dem zusätzlichen Tag 'Heur' aus, *XIP* vergibt 'GUESSED' als Feature, bei *MPT* erfolgt keine gesonderte Kennzeichnung. POS-Tags werden von *MPT*, *MS* und *XIP* auch an die Tokens vergeben, die nicht lemmatisiert sind. *TreeTagger* bezeichnet unbekannte Wortformen im Standardausgabeformat mit dem Lemma 'UNKNOWN' bzw. 'unknown'. Die POS-Tags der Wortformen, denen kein spezifisches Lemma zugewiesen werden kann, werden von *TreeTagger* anhand der Kontextinformationen und einem internen Suffixlexikon bestimmt. Die niedrige Anzahl der lemmatisierten Wortformen ist bei *TreeTagger* durch das Fehlen einer Morphologiekomponente bedingt, lemmatisiert werden nur Wortformen, die im internen Lexikon enthalten sind. *Glemm* vergibt an Wortformen, die anhand des Lexikons und der Morphologieregeln nicht zu analysieren sind, kein Lemma und kein POS-Tag. Der Prozentsatz der nicht erkannten Types zeigt die Auswirkungen der morphologischen Analyse auf die Erkennungsrate im Vergleich zu einem rein lexikonbasierten Verfahren. Beachtenswert ist der große Einfluss, den die Integration der Lemmata, die sich aus den nicht analysierbaren Wortformen konstituieren, bei *MPT*, *MS* und *XIP* auf die Länge der Lemmalisten hat. Sie verdeutlicht, dass es sich bei den nicht lemmatisierbaren Wortformen um Tokens handelt, die jeweils nur sehr selten im Korpus vorkommen.

Sehr unterschiedlich fallen auch die Häufigkeiten für Lemmata aus, die an bestimmten Positionen der frequenzsortierten NLP-Tool-spezifischen Lemmalisten stehen. Die Lemmata an Position 326.946, die die Länge der BLL 1.0 darstellt, haben abweichende Frequenzen. Das Lemma an dieser Position in der Lemmaliste von *glemm* (*Systemfußball*) kommt 73-mal im Korpus vor, das Lemma in der Lemmaliste von *MS* (*Hauptschuljahr*) ist 179-mal im Korpus vertreten (vgl. Tab. 5). Das erste Lemma mit einer Frequenz von 73 (*01069*) steht bei *MS* an Position 576.470, das erste Lemma mit der Frequenz von 179 (*abgabepflichtig*) bei *glemm* an Position 202.837.

Der Grund für die Abweichungen liegt im Lemmatisierungskonzept der NLP-Tools. In den Lemmalisten von *MPT*, *MS* und *XIP* sind im vorderen Bereich beispielsweise viele Tokens enthalten, die aus Zahlen und weiteren nicht-alphabetischen Zeichen bestehen (*41,2*, *730.000*). Diese Zeichenkombinationen werden von *TreeTagger* mit dem Lemma *@card@* ausgezeichnet, im Gegensatz zu Tokens, die nur aus Zahlen bestehen, und mit diesen als Lemma annotiert werden. In der Lemmaliste von *glemm* sind Lemmata mit nicht-alphabetischen Zeichen nicht vertreten. Auch nach der Bereinigung der Lemmalisten von Einträgen, die als Kandidaten für die BLL der nhd. Standardsprache aufgrund des Fehlens von alphabetischen Zeichen oder der Existenz von Sonderzeichen (#, @, |, \, /, …) nicht in Frage kommen, verbleiben große Unterschiede. In den Lemmalisten von *MPT*, *MS* und *XIP* stehen auch zahlreiche Kürzel (*Izt*, *ech*), die sich aufgrund ihrer Form durchaus als Lemmata eignen.

Die Abweichungen in den Lemmalisten von *glemm* und *TreeTagger*, die nur Tokens ein Lemma zuweisen, die durch das Lexikon oder die morphologische Analyse als gültige Wortformen identifiziert werden, sind geringer. Doch würde auch bei diesen beiden NLP-Tools die Angabe einer Mindestfrequenz für die Aufnahme der Lemmata in eine Lemmaliste zu unterschiedlich langen Lemmalisten führen. (Die Lemmaliste von *glemm* bis zu einer Frequenz von 70 hat 334.697 Einträge, die Lemmaliste von *TreeTagger* lediglich 308.311 Einträge.) Das Frequenzkriterium, das auf den ersten Blick als ein objektives, zahlenmäßig bestimmbares Maß erscheint, entpuppt sich auf den zweiten Blick als ein Kriterium, das primär die

den NLP-Tool-spezifischen zugrunde liegenden Lemmatisierungskonzepte widerspiegelt und nur sekundär einen Bezug zur Korpusgrundlage aufweist.

	glemm	*MPT*	*MS*	*TreeTagger*	*XIP*
unbereinigte Lemmalisten	73	177	179	74	177
bereinigte Lemmaliste	73	158	169	62	164

Tab. 5: Frequenz des Lemmas an Position 326.946

4.1.2 Die Verteilung von Wortformen auf Lemmaformen

Die automatische Zuweisung von Lemmata und POS-Tags an Wortformen im Korpus wird maßgeblich durch die internen Lexika und Morphologiekomponenten der NLP-Tools gesteuert. Der Kontext wird zur Auflösung von Ambiguitäten bei der Lemmatisierung oder dem POS-Tagging mit einbezogen (vgl. Abschnitt 4.1.6). Die Wortformen, die unter einer Lemmaform subsumiert sind, variieren mitunter zwischen den NLP-Tools. Eine Vereinheitlichung der Lemmaformen und der Lemmaparadigmen der NLP-Tool-spezifischen Lemmalisten steht zu Beginn der Erstellung der BLL. (In den folgenden Tabellen werden bei Bedarf die Wortformen, die einem Lemma von den NLP-Tools zugeordnet sind, unter dem Lemma oder der absoluten Frequenz gezeigt.)

Morphologische und orthographische Varianten

Die Wortformen von Lexemen, die in morphologischen Varianten vorkommen (*zerdeppern – zerteppern*), werden von den NLP-Tools meistens, aber nicht immer, derjenigen morphologischen Grundform zugeordnet, von der sie abstammen. Durch die Subsumption der flektierten Formen beider morphologischer Varianten unter nur einer Grundform ergeben die Frequenzen nicht unbedingt ein Bild des tatsächlichen Vorkommens der morphologischen Varianten im Sprachgebrauch, sondern sind in diesen Fällen beeinflusst durch das Lemmatisierungskonzept der NLP-Tools. Im folgenden Beispiel bildet *XIP* die flektierten Formen von *zapplig* und *zappelig* auf das Lemma *zappelig* ab.

Lemma	*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
zappelig ADJ	1.307	1.257	1.054	1.224	1.346
zapplig ADJ	94	96	93	82	–

Tab. 6: Zuordnung von Wortformen zu Lemmaformen – morphologische Varianten

Die Lemmatisierung orthographischer Varianten verhält sich ähnlich. Während in den meisten Fällen die Wortformen unter den jeweils orthographisch verwandten Lemmata subsumiert werden, gibt es immer wieder Abweichungen. *MPT* weist beispielsweise die flektierten Formen von *nummerieren* und *numerieren* dem Lemma *nummerieren* zu, und den flektierten Formen von *Tip* und *Tipp* das Lemma *Tipp*. *TreeTagger* hingegen subsumiert alle flektierten Formen von *Tip* und *Tipp* unter dem Lemma *Tip*. Daneben gibt es auch schwer nachzuvollziehende Entscheidungen bei der Zuordnung von Wortformen zu Lemmaformen, die nur einzelne flektierte Formen einer Flexionsreihe betreffen. Im Falle der Konjunktion *dass* beispielsweise, die von *TreeTagger* in ihrer groß geschriebenen Form zu Satzbeginn dem Lemma *daß* zugeordnet wird, während die klein geschriebene Form auf das Lemma *dass*

abgebildet wird. Auf einen Fehler im Lexikon zurückzuführen ist wohl auch, dass *TreeTagger* die gesamten flektierten Wortformen von *Alptraum* und *Albtraum* auf *Alptraum* abbildet, außer der Wortform *Albträumen*, die das Lemma *Albtraum* erhält.

Lemma	*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
nummerieren V	1.436	3.349	2.581	1.232	1.450
numerieren V	976	–	1.817	917	1.022
Tipp NN	159.540	150.176	65.429	–	94.758
Tip NN	84.159	–	51.679	139.518	47.399
dass CONJ	13.525.023 dass/Dass/daß/Daß	14.242.391 dass/Dass/daß/Daß	13.565.524 dass/Dass/daß/Daß	7.832.388 dass	8.387.070 dass/Dass
daß CONJ	–	–	–	6.416.781 Dass/daß/Daß	5.870.467 daß/Daß
Albtraum NN	15.649	13.609	13.372	1.309	12.791
Alptraum NN	21.028	18.035	17.056	30.023	16.880

Tab. 7: Zuordnung von Wortformen zu Lemmaformen – orthographische Varianten

In der Lemmastrecke von *elexiko* sind morphologische und orthographische Varianten über die Datenmodellierung miteinander verknüpft. Morphologische Varianten werden als Lemma in die WW-Stichwortliste übernommen, während orthographische Varianten auf die heutige Standardschreibweise abgebildet werden. In der BLL sind daher nur die heutigen Standardschreibweisen von Lexemen als Lemma enthalten. Die subsumierten orthographisch varianten Lemmata und ihre Frequenzen können in Kontrolldateien, die während eines Programmlaufs entstehen, eingesehen werden. Die Standardschreibweise entspricht nicht immer der tatsächlich häufigeren Realisation der Wortformen im Korpus.

Standardschreibweise		traditionelle Schreibweise			Standardschreibweise	
Soziografie	0	Soziographie	21	→	Soziografie	21
Spagetti	480	Spaghetti	7.310	→	Spagetti	7.790
Spagettiplausch	0	Spaghettiplausch	399	→	Spagettiplausch	399
Spagettiessen	0	Spaghettiessen	149	→	Spagettiessen	149
Spagettifresser	0	Spaghettifresser	38	→	Spagettifresser	38
Spagettiträger	0	Spaghettiträger	184	→	Spagettiträger	184
Spargelkremesuppe	0	Spargelcremesuppe	160	→	Spargelkremesuppe	160

Tab. 8: Subsumption orthographischer Varianten unter der Standardschreibweise (Lemmaliste *TreeTagger*)

Movierungen

Ein anderer Bereich, in dem die Lemmaformen häufig abweichen, sind Movierungen. In der WW-Stichwortliste werden Substantive, die zur Spezifizierung vom Genus meist weiblicher Personen- und Tierbezeichnungen oder zur Spezifizierung von Pluralformen aus den entsprechenden männlichen Wörtern gebildet werden, mit allen Grundformen als Lemma genannt (*Löwe – Löwin*, *Rechtsanwalt – Rechtsanwältin*, *Minister – Ministerin*). Auch die NLP-Tools führen in der Regel die weibliche und männliche Form als Lemma und subsumieren unter ihnen die entsprechenden Wortformen. Lediglich bei Nomen, die mit dem Gebrauch des bestimmten oder unbestimmten Artikels zwischen starker und schwacher Flexion wechseln (*Beamte – Beamter*), variiert die maskuline Lemmaform: Während jedes NLP-Tool die *Beamtin* als Lemma liefert, subsumieren *MPT* und *XIP* die männlichen Wortformen und Pluralwortformen unter dem Lemma *Beamte*, *glemm* und *TreeTagger* unter dem Lemma

Beamter. Werden dieselben flektierten Wortformen von den NLP-Tools unterschiedlichen Lemmaformen (*Beamte – Beamter*) zugeordnet, kommt es in der BLL zu einer Dopplung dieser Frequenzen.

Lemma	*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
Beamt NN Beamte/Beamten	–	–	358.471	–	–
Beamte NN Beamte/Beamter/Beamten	–	403.034	–	–	394.414
Beamter NN	429.455 Beamte/Beamter/ Beamten	–	31.201 Beamter	396.683 Beamte/Beamter/ Beamten	–
Beamtin NN Beamtin/Beamtinnen	9.898	9.128	8.847	8.953	8.027

Tab. 9: Zuordnung von Wortformen zu Lemmaformen – Movierungen

Movierungen werden bei der Erstellung der BLL nicht mit den meist männlichen Grundformen unifiziert, sondern bleiben als eigener Eintrag erhalten.

Artikel und Pronomen

Die folgende Tabelle, die die Verteilung der flektierten Wortformen des bestimmten und unbestimmten Artikels sowie eines Pronomens auf die Lemmaform(en) darstellt, verdeutlicht zum einen, dass Lemmaformen in abweichender Granularität für das gleiche Wortformenparadigma angesetzt werden. Zum anderen wird ersichtlich, dass bei gleicher Granularität der Lemmaformen die Wortformen, die unter ihnen zusammengeführt werden, variieren. Auch die Lemmaformen, die das gleiche Wortformenparadigma repräsentieren, sind mitunter different, während identische Lemmaformen unterschiedliche Wortformen subsumieren.

Lemma BLL	*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
der,die,das ART	d- 372.041.251 der/die/das/den/ dem/des	die 195.144.842 die/der/den der 140.798.904 der/den/des/dem das 63.511.136 das/des/dem	die 213.676.672 die/der/den der 128.782.038 der/den/des/dem das 64.028.320 das/des/dem	die 348.613.119 der/die/das/den/ dem/des	der 213.132.400 der/den/dem/des die 116.675.041 die das 29.662.783 das
ein(e) ART	ein 21.870.507 eine/ein/einen/ einem/einer/eines	ein 42.900.667 ein/einen/einem/ eines eine 29.293.374 eine/einer	ein 72.760.859 eine/ein/einen/ einem/einer/eines	eine 75.238.859 eine/ein/einen/ einem/einer/eines	ein 77.695.214 eine/ein/einen/ einem/einer/eines

dies(e,er,es) PRON	dies 14.815.935 dieser/dies/dieses/ diese/diesem/diesen	dieser 15.421.438 dieser/dies/dieses/ diese/diesem/diesen	dieser 15.009.815 dieser/dies/dieses/ diese/diesem/diesen	diese 14.064.408 dieser/Dies/dieses/ diese/diesem/diesen dies 1.265.544 dies	dies 3.579.915 dieser/dies/dieses/ diese/diesem/diesen dies (DET) 12.260.678 dieser/dieses/diese/ diesem/diesen

Tab. 10: Zuordnung von Wortformen zu Lemmaformen – Artikel und Pronomen

Für den Vergleich der NLP-Tool-spezifischen Lemmalisten untereinander und mit der WW-Stichwortliste werden die unterschiedlichen Lemmaformen von Artikeln, Pronomen und (seltener) Adjektiven auf eine Lemmaform abgebildet, die die gesamte Flexion des Lemmas enthält. Die Vereinheitlichung der verschiedenen Lemmaformen wird über händisch erstellte Listen vorgenommen, die in den Programmlauf eingebunden sind. Werden mehrere Lemmata in einer Lemmaliste auf die neue Lemmaform abgebildet, werden die Frequenzen addiert. Das POS-Tag bildet auch hier ein distinktives Merkmal, unifiziert werden nur Lemmata mit demselben POS-Tag.

Partizipien und Substantivierungen

Kriterien für die Aufnahme von Partizipien in die Stichwortlisten von Wörterbüchern werden in Erb (2005: 93f.) anhand des Vorgehens bei *elexiko* eingehend erläutert. Im Allgemeinen werden Partizipien in Wörterbüchern unter dem Lemma des Verbs subsumiert, wenn sie keine lexikalisierte Bedeutung haben, die sich von der des Verbs unterscheidet. Die NLP-Tools verwenden Partizipien und Substantivierungen fast immer als zusätzliches Lemma neben der verbalen Grundform, unabhängig von Bedeutungsgleichheit oder Lexikalisierung. (Lexikalisierte Adjektive werden in den folgenden Tabellen fett markiert.)

Lemma		*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
ansprechen	V	148.696	150.422	129.919	143.302	145.693
ansprechen	ADJ	–	–	–	–	44.876
ansprechend	ADJ	34.435	26.847	28.523	33.248	–
angesprochen	ADJ	13.333	15.909	18.979	13.256	–
Ansprechen	NN	–	834	1.108	957	3.810
Ansprechen	V	28	–	–	–	–
aussprechen	V	51.544	205.241	187.366	177.860	155.870
aussprechen	ADJ	–	–	–	–	51.617
aussprechend	ADJ	–	88	108	2	–
ausgesprochen	ADJ	176.227	29.050	28.135	51.672	27.234
Aussprechen	NN	–	740	828	735	838
Aussprechen	V	23	–	–	–	–
laufen	V	1.025.978	1.025.292	962.325	983.085	988.266
laufen	ADJ	–	–	30	–	303.309
laufend	ADJ	299.104	275.817	270.384	292.893	–
gelaufen	ADJ	5.079	6.232	10.956	5.610	–
Laufen	NN	46.191	505	41.356	38.903	47.152
Laufen	NE	–	25.217	–	661	–

gehen	V	5.350.943	5.233.632	5.160.944	5.231.625	5.273.193
gehen	ADJ	–	–	64	–	45.411
gehend	ADJ	–	23.747	23.895	27.793	–
gegangen	ADJ	15.835	16.273	15.330	5,755	–
Gehen	NN	28.353	16.689	22.626	5.704	17.671

Tab. 11: Lemmatisierung von Partizipien und Substantivierungen

Die Wortformen, die unter einem Lemma zusammengefasst werden, erklären die mitunter stark abweichenden Frequenzen der Verben, Partizipien und Substantivierungen zwischen den NLP-Tools. *XIP* subsumiert substantivierte Partizip-2-Formen (*Angesprochene*) nicht wie die anderen NLP-Tools unter der adjektivischen Lemmaform *angesprochen*, sondern unter dem nominalen Lemma *Ansprechen*; daher sind die Frequenzen der Substantivierungen bei *XIP* höher. *XIP* vergibt im Gegensatz zu den weiteren NLP-Tools für Partizipien als Lemmaform fast immer die Infinitivform des Verbs, die sich durch das POS-Tag ADJ von der verbalen Lemmaform unterscheidet.

Glemm subsumiert unter den nominalen Lemmaformen *Ansprechen* und *Aussprechen* lediglich die nominalen Wortformen im Genitiv (*Ansprechens*, *Aussprechens*). Die nominalen Wortformen *Ansprechen* und *Aussprechen* werden im Gegensatz zu den anderen NLP-Tools mit der verbalen Lemmaform (*ansprechen*, *aussprechen*) lemmatisiert. Die Wortform *ausgesprochen* wird von *glemm* immer dem adjektivischen Lemma *ausgesprochen* zugeordnet, während die anderen NLP-Tools den Kontext und damit die verbale oder adjektivische Verwendung der Wortform im Satz in ihre Analyse mit einbeziehen. Die Wortform *ausgesprochen* wird je nach Verwendung von den Taggern und Parsern dem Lemma *aussprechen* oder dem Lemma *ausgesprochen* zugewiesen. Die Zahlen in der obigen Tabelle lassen erkennen, dass die verbale Verwendung häufiger ist. *Glemm* verzichtet häufig auf die Nennung einer Lemmaform, die dem Partizip 1 entspricht, und subsumiert die Partizip-1-Formen, die im Korpus auftreten, unter der verbalen Lemmaform. Dadurch werden zwar die Lemmaformen reduziert, die für ein verbales Wortformenparadigma anzusetzen sind, doch sind auch viele lexikalisierte Partizip-1-Formen wie *entscheidend*, *kommend* oder *rauschend* in der NLP-Tool-spezifischen Lemmaliste von *glemm* nicht zu finden.

Während der Erstellung der BLL werden Partizipien und Substantivierungen auf ihre Grundformen abgebildet, wenn die Partizipien und Substantivierungen nicht als eigener Eintrag in der WW-Stichwortliste vorhanden sind. Alignierungsalgorithmen führen Partizipien und Substantivierungen, die sich nicht in der WW-Stichwortliste befinden, auf ihre entsprechenden verbalen Grundformen zurück und addieren deren Frequenz zur Häufigkeitsangabe der Grundform (vgl. Anhang). Auf diese Weise wird die tatsächliche Frequenz der Lemmata angegeben, inklusive der Häufigkeiten der nicht lexikalisierten Wortformen, die der Flexionsreihe der Grundform angehören. Die zahlreichen Partizipien und Substantivierungen ohne Bedeutungsunterscheidung zur Grundform werden in der BLL nicht verzeichnet.[29]

[29] Sollte es eine spezifische Fragestellung erfordern, können Partizipien und Substantivierungen auch anhand der Lemmata in der WW-Stichwortliste verifiziert und als zusätzlicher Eintrag in die BLL übernommen werden, anstatt die Frequenzen zu den Grundformen zu summieren.

	glemm	*MPT*	*MS*	*TreeTagger*	*XIP*
Partizip 1	229	8.313	11.197	4.086	–
Partizip 2	2.302	5.865	5.726	2.714	316
Substantivierung	7.055	5.212	12.369	5.826	12.412

Tab. 12: Anzahl der alignierten Partizipien und Substantivierungen

Zusätzlich zu den Einträgen in der WW-Stichwortliste werden in die BLL adjektivisch gebrauchte Partizipien und Substantivierungen aufgenommen, die häufiger vorkommen als ihre Grundformen. Damit soll dem tatsächlichen Sprachgebrauch Rechnung getragen werden, die Partizipien und Substantivierungen unterscheiden sich semantisch nicht von der Grundform. Ihr häufiges Auftreten im Korpus wird als relevant für das Vorkommen in der BLL angesehen, da die Partizipien oder Substantivierungen gebräuchlicher sind als die unter der Grundform zusammengefassten Wortformen. Betroffen sind ca. 600 Partizipien und Substantivierungen, die zusätzlich in der BLL stehen.

Basis			Konversion		
segelfliegen	V	0	Segelfliegen	NN	1.212
windsurfen	V	0	Windsurfen	NN	763
ausparken	V	598	Ausparken	NN	3.147
aufbrausen	V	496	aufbrausend	ADJ	2.668
beängstigen	V	465	beängstigend	ADJ	10.543
stressen	V	2.611	gestresst	ADJ	4.611

Tab. 13: Verbale Grundformen, Substantivierungen und Partizipien (Lemmaliste *TreeTagger*)

4.1.3 Übergeneralisierung

Es existieren zahlreiche Lemmaformen im Deutschen, die sich von der üblichen Zitierform für Lemmata unterscheiden. Diese Lemmaformen weichen vom lexikografischen Usus zur Bildung einer Lemmaform durch die Reduktion der Wortformen auf eine Grundform im Singular oder Positiv ab. Betroffen von der falschen Rückführung der Wortbestandteile sind nicht nur Wortendungen, sondern auch flektierte lexikalische Morpheme zu Wortbeginn oder im Wortinneren. Die korrekten Lemmaformen können nur über Listen bestimmt werden, die die Ausnahmen von den lexikografischen Konventionen verzeichnen. Viele Irregularitäten der deutschen Lexik sind in Rechtschreibkorrekturprogrammen integriert, regelmäßig gebildete Wortformen werden als Fehler angezeigt (*Flitterwoche*, *Minderlohn*, *nächstgut*). Von den NLP-Tools wird die richtige Lemmatisierung bestimmter Wortreihen weitgehend ignoriert.

Pluraliatantum

Pluraliatantum und überwiegend im Plural gebräuchliche Nomen werden bei der Lemmatisierung analog zu den allgemeinen Lemmatisierungsregeln auf eine vermeintliche Lemmaform im Singular zurückgeführt, die im Deutschen nicht üblich ist. Eine pluralische Lemmaform für Pluraliatantum ist in den NLP-Tool-spezifischen Lemmalisten nur selten vorhanden. Die allgemeine Rückführung von Wortformen im Plural auf Lemmaformen im Singular wird von allen NLP-Tools gleichermaßen praktiziert. In den folgenden Beispielen liegen im Korpus mindestens 98% der Wortformen im Plural vor.

Lemma		*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
Flitterwoche	NN	3.405	3.152	56	3.099	3.130
Flitterwochen	NN	–	–	3.037	–	–
Spielsache	NN	16.815	16.210	15.757	16.163	16.169
Wintersache	NN	543	470	475	471	474
Englischkenntnis	NN	3.059	3.076	3.073	3.106	3.109
Kontodatum	NN	1.059	1.192	1.178	1.146	1.307
Lebensumstand	NN	8.293	8.212	8.049	8.227	8.234
Kokosflocke	NN	132	–	138	135	21
Kokosflocken	NN	–	43	–	–	116
Kokosflocken	NE	–	93	–	–	–

Tab. 14: Lemmatisierung von Pluraliatantum

Die *Kokosflocken* werden von *MPT* nicht erkannt und daher die Wortform als Lemma verwendet. Die kontextuellen Bedingungen, die zur Erkennung der *Kokosflocken* als Eigen- oder Gattungsnamen führen, sind nicht nachzuvollziehen. *XIP* nimmt einen unmittelbar vorausgehenden bestimmten Artikel als Indiz, um die *Kokosflocken* auf eine Lemmaform im Singular zu lemmatisieren, während die Pluralformen, die ohne Artikel stehen, auf die Lemmaform im Plural abgebildet werden.

Alignierungsalgorithmen übernehmen den Abgleich der Lemmaformen im Singular in den NLP-Tool-spezifischen Lemmalisten auf die in der WW-Stichwortliste verzeichneten korrekten Pluralformen bei der Erstellung der BLL (vgl. Anhang). Die lexikografischen Informationen, die in der WW-Stichwortliste durch die Lemmaformen im Plural enthalten sind, werden benötigt, um in der BLL korrekte Lemmaformen zur Verfügung zu stellen. Eine allgemeine Erweiterung bestimmter Kompositaendbestandteile auf die Pluralform ist nicht möglich, es gibt mit den gleichen Kompositaendbestandteilen auch Komposita im Singular (*Hochzeitsdatum*, *Menschenkenntnis*, *Ehrensache*, *Kehrwoche*).

Ambige nominale Pluralwortformen

Einige pluralische Lemmaformen haben singularische Pendants, die sich in ihrer Bedeutung von den Lemmata der Pluraliatantum unterscheiden. Während in der WW-Stichwortliste beide Lemmaformen enthalten sind (*Anstalt*/*Anstalten*, *Unterlage*/*Unterlagen*), subsumieren die NLP-Tools die Pluralwortformen, die zum Flexionsparadigma beider Lemmaformen gehören können, immer unter einer Lemmaform im Singular. Die Frequenzen der Pluralwortformen fließen in die BLL durch die Lemmaform im Singular ein, da die NLP-Tool-spezifischen Lemmalisten ein Lemma im Plural nicht anbieten.

Lemma		*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*	*Wortform*	
Unterlage	NN	85.871	81.565	75.511	78.074	78.164	Unterlage	~ 4.700
Unterlagen	NN	–	–	–	–	–	Unterlagen	~ 81.100
Anstalt	NN	44.987	43.715	38.566	39.541	39.562	Anstalt	~ 26.900
Anstalten	NN	–	–	–	–	–	Anstalten	~ 17.800

Tab. 15: Lemmatisierung ambiger nominaler Pluralwortformen

Pluralische Erstglieder in Nomenkomposita

Pluralmorpheme kommen nicht nur als Endung eines Nomenkompositums vor, sondern bilden auch eine Komponente zu Beginn oder innerhalb einer nominalen Wortform im Singular. In diesen Fällen variiert die Zuweisung einer korrekten Lemmaform zwischen den NLP-Tools.

Lemma		*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
Kraftverhältnis	NN	61	7.681	74	54	7.664
Kräfteverhältnis	NN	7.610	–	7.413	7.602	–
Landspiel	NN	–	49.127	44.817	–	–
Länderspiel	NN	61.101	–	–	39.048	39.048
Spielsachebasar	NN	–	220	–	–	190
Spielsachenbasar	NN	201	–	178	188	–
Blumeregen	NN	–	–	–	–	80
Blumenregen	NN	81	81	52	76	–
Wellengang	NN	1.802	1.752	1.697	1.758	1.762

Tab. 16: Pluralische Erstglieder in Nomenkomposita

Superlative in Komposita

Ähnlich wie mit den Pluralmorphemen in Nomenkomposita, die fälschlicherweise auf eine Singularform abgebildet werden, verhält es sich mit Steigerungsformen in zusammengesetzten Nomen und Adjektiven. Die Rückführung des Morphems im Komparativ oder Superlativ auf den Positiv führt zu einer Lemmaform, die es im Deutschen nicht gibt. Bei Adjektiven wird die Rückführung von Superlativen auf Positive von *glemm*, *MS*, *TreeTagger* und *XIP* durchgeführt, während *MPT* die Steigerung in der Lemmaform beibehält. Bei Nomenkomposita mit Superlativ wird hingegen die korrekte Lemmaform von *glemm*, *MS*, *TreeTagger* und *XIP* angegeben, während *MPT* das Morphem im Superlativ auf den Komparativ reduziert.

Lemma		*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
nahgut	ADJ	–	–	–	–	1.578
nächstgut	ADJ	1.599	–	1.565	1.560	–
nächstbesten	ADJ	–	887	–	–	–
nächstbe	ADJ	–	544	–	–	–
weltgut	ADJ	10.143	–	9.350	9.420	9.660
weltbesten	ADJ	–	6.652	–	–	–
weltbe	ADJ	–	1.915	–	–	–
weltgroß	ADJ	30.726	–	–	29.966	–
weltgrößten	ADJ	–	13.310	–	–	30.551
Weltgrößte	NN	–	13.732	–	–	–
Minderlohn	NN	–	37.407	–	–	–
Mindestlohn	NN	43.986	–	34.758	37.260	27.433

Tab. 17: Superlative in Adjektivkomposita

MPT stellt für die Adjektivformen im Superlativ zwei Lemmaformen bereit. Unter unvollständigen adjektivischen Lemmaformen (*nächstbe*, *weltbe*) oder einer weiblichen nominalen Lemmaform (*Weltgrößte*) werden die weiblichen adjektivischen Wortformen im Nominativ und Akkusativ zusammengefasst. Die vollständigen adjektivischen Lemmaformen (*nächstbesten*, *weltbesten*, *weltgrößten*) subsumieren bei *MPT* alle weiteren flektierten Formen der Adjektive im Superlativ. *MS* lemmatisiert viele Wortformen, die ein ‘ß’ enthalten nicht, auch die flektierten Formen von *weltgrößte*. Die Frequenzen der Wortformen werden in der Tabelle nicht aufgeführt. Adjektivkomposita, in denen der Superlativ zu Wortanfang steht, werden von allen NLP-Tools richtig erkannt (*bestmöglich*, *bestplatziert*, *größtmöglich*).

4.1.4 Diskontinuierliche Konstituenten

Die im Satz auseinanderstehenden Bestandteile von Präverbfügungen werden von den NLP-Tools unterschiedlich behandelt. *MPT* und *XIP* weisen den Präfixen der Präverbfügungen das POS-Tag PREP oder ADV zu, *glemm*, *MS* und *TreeTagger* vergeben als eigenes POS-Tag für Verbzusätze VRZ, VPART bzw. PTKVZ und unterscheiden damit für die betroffenen Wortformen ihre Funktion als verbales Präfix von der als Präposition oder Adverb. Die abgetrennten Verbformen hingegen werden von allen NLP-Tools wie eigenständige Verben mit dem POS-Tag V annotiert, als Lemma wird für die Verbform eine Lemmaform ohne das Präfix verzeichnet.

Verb mit Präfix	Lemma	*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
(*ab*/*aus*)*statten*	statten	14.534	29.860	192.458	14.409	14.551
(*bei*)*pflichten*	pflichten	10.074	9.697	9.514	9.502	9.651
(*ein*)*heimsen*	heimsen	8.000	7.218	6.944	7.105	7.204
(*an*)*freunden*	freunden	2.251	5.169	6.320	2.999	5.317
(*auf*)*heitern*	heitern	–	1.089	3.491	1.008	1.130

Tab. 18: Verbale Komponenten von Präverbfügungen

Die Zuweisung von Lemmata, die das Präfix nicht enthalten, an die unvollständigen verbalen Komponenten der Präverbfügungen, führt zum einen zum Auftreten von Verben in den NLP-Tool-spezifischen Lemmalisten, die es im Deutschen nicht gibt. Diese im Deutschen nicht existenten Lemmata der Lemmalisten werden durch den Abgleich mit der WW-Stichwortliste aussortiert. Zum anderen werden aber die Frequenzen der Lemmata der Verben erhöht, die es im Deutschen auch ohne Präfix gibt ((*ein*)*schlafen*, (*hinaus*)*gehen*, (*auf*)*bieten*). Diese Frequenzen fließen fälschlicherweise mit den Lemmata ohne Präfix in die BLL ein.

Die zahlenmäßig großen Abweichungen des Lemmas *statten* in den Lemmalisten von *MPT* und *MS* liegen an den Wortformen, die im Gegensatz zu den anderen NLP-Tools unter dem Lemma subsumiert werden. *MPT* ordnet unter dem Lemma *statten* fälschlicherweise das Partizip *gestattet* ein, *MS* reiht den ersten Bestandteil des Verbs *stattfinden* in getrennter Wortstellung in das Flexionsparadigma von *statten* mit ein. Die höheren Frequenzen von *freunden* bei *MPT*, *MS* und *XIP* sind darauf zurückzuführen, dass die drei Tools im Gegensatz zu *glemm* und *TreeTagger* auch die nominalen Wortformen *Freunde* und *Freunden* unter dem Lemma subsumieren.

Beim Vergleich der häufigsten Verbzusätze fallen zwischen *MS* und *TreeTagger* Diskrepanzen in den Frequenzen auf. In der Lemmaliste von *glemm* hingegen fehlen die häufigsten Verbzusätze von *MS* und *TreeTagger*. *Glemm* zeichnet nur diejenigen verbalen Präfixe als solche aus, die nicht auch als Adverb oder Präposition existieren, wenn sie alleine stehen. Für eine Disambiguierung der POS-Tags ist der Kontext relevant, der von *glemm* in die Analyse nicht mit einbezogen wird. Im Falle der Tagger und Parser wäre es vorzuziehen, dass die erkannten Präfixe zusammen mit der verbalen Komponente zusätzlich in die Lemmaform des Verbs einfließen.

glemm		*MS*		*TreeTagger*	
hervor	193386	an	2863431	an	2830000
dar	158624	aus	2387237	aus	2580428
teil	158554	ein	2142518	ein	2154082
halt	112073	auf	1862949	ab	2000534
raus	105221	ab	1463348	auf	1786985
drauf	89358	vor	1426394	vor	1617144
heran	67301	zu	1088488	zurück	1176699
ran	44256	mit	1048286	statt	962732
runter	34961	zurück	1037089	mit	846588
inne	31229	statt	635049	zu	828846
herunter	30628	hin	473390	zusammen	537539
heim	28052	nach	433392	fest	473711
beiseite	23208	her	255611	hin	439320
durcheinander	22825	bei	233315	weiter	417921
…		...		...	

Tab. 19: Frequenzsortierte Verbalpräfixe

4.1.5 Fehlende Lemmata in den NLP-Tool-spezifischen Lemmalisten

Durch einen Vergleich der NLP-Tool-spezifischen Lemmalisten mit der *elexiko*-Stichwortliste kann festgestellt werden, welche Lemmata der nhd. Standardsprache in den NLP-Tool-spezifischen Lemmalisten fehlen. Die *elexiko*-Stichwortliste enthält nach der Unifikation von Artikeln und Pronomen auf eine Form, die diese mit den Lemmalisten vergleichbar macht (vgl. Abschnitt 4.1.2), 304.199 Lemmata. Für den Vergleich mit den Lemmalisten werden die Mehrwortlemmata nicht berücksichtigt, die keine zusammengeschriebene Variante besitzen (*allgemein verträglich*, *bunt gemustert*, *voll getankt*, *wohl riechen*, *wund gerieben*, *Agnus Dei*), wodurch 302.828 Lemmata in der Lemmastrecke von *elexiko* verbleiben. Die Anzahl der Lemmata von *elexiko*, die nicht in den NLP-Tool-spezifischen Lemmalisten enthalten sind, variiert je nach NLP-Tool zwischen 10.411 (*MS*) und 97.395 (*XIP*).

	glemm	*MPT*	*MS*	*TreeTagger*	*XIP*
elexiko-Stichwortliste	3,8%	8,4%	3,4%	17,1%	32,1%

Tab. 20: Prozentsatz der *elexiko*-Lemmata, die nicht in den NLP-Tool-spezifischen Lemmalisten enthalten sind

Aufgrund der unterschiedlichen Lemmatisierungskonzepte der NLP-Tools sind die Lemmata der *elexiko*-Stichwortliste, die in den Lemmalisten der einzelnen NLP-Tools nicht gefunden werden, sehr heterogen. Da *MPT*, *MS* und *Xerox* Wortformen, die nicht analysiert werden können, die Wortform selbst als Lemma sowie ein POS-Tag zuweisen, enthalten deren Lemmalisten neben einer großen Anzahl an Nichtwörtern (*ÿZiÿzek*, *üngstwe*, *zwùanzg*, *zwöls*) auch seltene und neue Fremdwörter (*Actionpainting*, *Adduktor*, *Aktinolith*, *Algebraiker*). *Glemm* und *TreeTagger* hingegen vergeben an Wortformen, die sie nicht erkennen, kein Lemma. Die meisten *elexiko*-Lemmata, die in der Lemmaliste von *glemm* fehlen, basieren auf fremd- bzw. umgangssprachlichen Morphemen (*Bierdimpfel*, *Bignonie*, *Epheserbrief*, *fiktionalisieren*) oder sind Eigennamen (*Minerva*, *Mombasa*, *Narbonne*). Die Lemmata, die in der Lemmaliste von *TreeTagger* nicht gefunden werden, stammen aus denselben Kategorien, hinzu kommen zahlreiche Komposita (*Ballspielerin*, *Basiswaschmittel, Bodensystem*, *Brustkorbatmung*). Die einzelnen Bestandteile der Komposita werden von *TreeTagger* erkannt, *TreeTagger* fehlt jedoch eine Morphologiekomponente, die Komposita zerlegen kann, und

lemmatisiert nur Wörter, die im internen Tagger-Lexikon stehen. Unbekannten Wortformen wird von *TreeTagger* anhand ihrer Endung und ihres Kontextes zwar ein POS-Tag zugewiesen, als Lemma jedoch wird „UNKNOWN“ angegeben (vgl. Abschnitt 4.1.1).

Durch die Rückführung einzelner Morpheme der Wortformen auf ihre Grundformen und die Zusammensetzung der Lemmaform aus den Grundformen der einzelnen Morpheme befinden sich in den Lemmalisten von *MPT* und *XIP* zahlreiche Lemmata, die sich von der gebräuchlichen Form von Komposita unterscheiden. Mit regelbasierten Algorithmen und unter Einbeziehung der auf das Lemma abgebildeten Wortformen können ca. 85% der Komposita aus der Lemmaliste von *MPT*, die aus Morphemgrundformen bestehen, auf die in der Lexikografie üblichen Lemmaformen der *elexiko*-Stichwortliste abgebildet werden. (Dieser Prozentsatz wäre durch zusätzlichen Aufwand bei der Implementierung der Algorithmen sicher noch zu steigern.) Ein großer Teil der Lemmata von *elexiko*, die in der Lemmaliste von *MPT* nicht gefunden werden, sind Komposita, deren Wortbildungsmuster durch die Alignierungsregeln nicht abgedeckt werden (*baum wolle feld* ‘Baumwollfeld’, *ein jahr vertrag* ‘Einjahresvertrag’, *arzt streik* ‘Ärtzestreik’, *bad wanne rennen* ‘Badewannenrennen’).

Bei *XIP* werden die Bestandteile von Komposita häufig auf eine verbale Grundform zurückgeführt. In welchen Fällen Morpheme auf eine verbale Grundform abgebildet werden und in welchen Fällen die Morphemgrundform ein Nomen oder Adjektiv darstellt, ist nicht nachzuvollziehen (*dunkeln#rot*, *hellen#rot*, *grell#rot*). Häufig stehen in der Lemmaliste von *XIP* mehrere Varianten für eine in der Lexikografie übliche Lemmaform zur Verfügung, in denen die einzelnen Wortbestandteile auf unterschiedliche Grundformen zurückgeführt sind. Wortformen, die zum Flexionsparadigma eines Lemmas gehören, werden häufig auf mehrere Lemmaformen verteilt. Die Lemmaform *backen#mitteln* wird z.B. für die Wortform *Backmitteln* gewählt, die Lemmaform *backen#Mittel* repräsentiert die Wortformen *Backmittel* und *Backmittels*. Die *Mangelwirtschaft* wird auf das Lemma *mangeln#Wirtschaft* abgebildet, außer der *Mangelwirtschaft* geht ein Substantiv mit einem ausgesparten Wortteil wie *Kommando-* voraus, in diesem Fall wird die Wortform dem Lemma *mangeln#wirten#Schaft* zugewiesen. Auch die unterschiedliche Behandlung von Fugenelementen äußert sich in differenten Lemmaformen, die für eine in der Lexikografie übliche Lemmaform stehen (*überleben#Bericht*, *überleben#berichten*, *überleben/s#Bericht*). Weil die linguistischen Annotationen von *XIP* im Erstellungszeitraum der BLL 1.0 nicht zur Verfügung standen, wird auf eine Lemmareparatur verzichtet und es werden lediglich die Sonderzeichen (#/}=), die die Morphemgrenzen kennzeichnen, entfernt und die Wortbestandteile zusammengefügt.

In der Lemmaliste von *MS* stehen zahlreiche Komposita mit einer falschen Lemmaform, weil einzelne Bestandteile nicht korrekt lemmatisiert werden. Während z.B. das einfache Nomen *Basis* als Lemma verzeichnet ist, enthalten daraus gebildete Komposita *Basisis* als Bestandteil (*Basisispunkt*, *basisisdemokratisch*). Aus dem Nomen *Brut* wird in Komposita *Brüt* (*Brütstätte*, *Brütgebiet*). Ein großer Teil der weiteren Lemmata der *elexiko*-Stichwortliste, die in der Lemmaliste von *MS* kein Gegenstück finden, besteht aus Nomenkomposita, deren vorderes Wort im Plural steht, das von *MS* auf eine Singularform zurückgeführt wird (*Buchberg*, *Krauthexe*, *Datumträger*, *Datumschutzbeauftragte*).

Insgesamt 369 Lemmata aus der WW-Stichwortliste haben kein Pendant in einer der NLP-Tool-spezifischen Lemmalisten. Eine Überprüfung dieser Lemmata ergibt, dass es sich überwiegend um Lemmata handelt, die entweder in der *elexiko*-Stichwortliste zu streichen sind, weil es sich bei den Lemmata lediglich um Wortformen des Wortformenparadigmas eines anderen Lemmas handelt (*blies*, *befahl*, *befiehl*), um fehlerhafte Lemmaformen (*Anschaf-*

fungskost) oder um Pluralia, die fälschlicherweise auch als Singularform in der Lemmastrecke stehen (*Atemwege*, *Gründerjahre*).[30] Nach einer sukzessiven Bereinigung verbleiben einige wenige Lemmata aus *elexiko*, die in keiner NLP-Tool-spezifischen Lemmaliste eine Übereinstimmung finden. Dies sind Morphologievarianten, deren zugehörige Wortformen von den NLP-Tools einem anderen Lemma zugeordnet werden (*sukzessive*, *rigide*), heute sehr ungebräuchliche Wörter (*personaliter*) und dialektal markierte Lexeme (*rott*). Alle weiteren Lemmata der *elexiko*-Stichwortliste werden in mindestens einer NLP-Tool-spezifischen Lemmaliste gefunden.

4.1.6 Ambige Wortformen

Unter ambigen Wortformen werden hier diejenigen textuellen Bausteine verstanden, denen in einem Kontext zwei Lemmaformen zugewiesen werden. Dass derselben Wortform in unterschiedlichen Kontexten unterschiedliche Lemmata und POS-Tags zugewiesen werden, ist bei NLP-Tools, die die sprachliche Umgebung in ihre Analysen mit einbeziehen, die Regel. Beispielsweise werden Partizip-2-Formen in den zusammengesetzten Zeiten mit ihrer verbalen Grundform lemmatisiert und erhalten das POS-Tag V, während, wenn sie als Beiwort von Substantiven dienen, die Partizip-2-Form die Lemmaform bildet, begleitet von dem POS-Tag ADJ.

Innerhalb desselben Kontextes vergeben nur *MPT* und *TreeTagger* zwei mögliche Lemmaformen. *TreeTagger* macht die Zuweisung von zwei unterschiedlichen Lemmata abhängig von der Vergabe abweichender POS-Tags an die Wortform und damit von unterschiedlichen syntaktischen Interpretationsmöglichkeiten, während *MPT* auch innerhalb einer POS-Kategorie zwei Lemmata vergibt. Betroffen von kontextuellen Mehrdeutigkeiten sind bei *MPT* 7.888 und bei *TreeTagger* 400.159 unterschiedliche Wortformen. Die hohe Anzahl der ambigen Wortformen bei *TreeTagger* ist auch auf die feine Granularität des Tagsets zurückzuführen, die ambigen Wortformen wurden mit dem Original-Tagset vor der Unifizierung der Tagsets ermittelt (vgl. Abschnitt 4.2.1). Beispielsweise erhält die Wortform *wenig* als attribuierendes Indefinitpronomen das Lemma *wenig*, als substituierendes Indefinitpronomen das Lemma *wenige*. Eine Gegenüberstellung der frequenzsortierten Listen mit ambigen Wortformen zeigt, dass *MPT* und *TreeTagger* ganz anderen Wortformen mehrere Lemmata zuweisen. In die Frequenzangaben der Lemmata in der BLL gehen die alternativen Lemmaformen und ihre POS-Tags von *MPT* mit einer Frequenz von 0,5 ein, TreeTagger berechnet genaue Wahrscheinlichkeiten für die Alternativen, welche in die Zählung einfließen.

[30] Die falsche Singularform in der *elexiko*-Stichwortliste korrespondiert mit der falschen Lemmatisierung von Pluralia durch die NLP-Tools. Dadurch können die später angewandten Alignierungsalgorithmen, die die falsche Singularformen der NLP-Tool-spezifischen Lemmalisten auf die Pluralform in *elexiko* abbilden, nicht mehr greifen. Das Lemma wurde bereits durch einen einfachen Stringvergleich einem falschen Lemma der *elexiko*-Stichwortliste zugeordnet.

MPT				*TreeTagger*			
der	die DET	der PRON	604271	ein	eine ART	ein PTKVZ	1452290
eins	einer PRON	eins NUM	134915	weiter	weit ADJD	weiter ADV	1028392
gehört	hören V	gehören V	44950	anderen	ander ADJA	andere PIS	617056
dem	das PRON	der DET	43369	wenig	wenige PIS	wenig PIAT	608639
den	die DET	der PRON	38294	bekannt	bekannt ADJD	bekennen VVPP	593764
kostet	kosen V	kosten V	35118	weiter	weit ADJD	weiter PTKVZ	565702
kosten	kosen V	kosten V	35022	weniger	wenige PIAT	weniger ADV	560594
Eins	einer PRON	eins NUM	23934	weniger	wenige PIS	weniger ADV	533296
gebracht	bringen V	gebrechen V	19321	andere	ander ADJA	andere PIS	501081
Walter	walter N	Walter NE	16952	früher	früh ADJD	früher ADV	464087
schienen	scheinen V	schienen V	16331	wenig	wenig ADV	wenige PIS	425889
gelangen	gelingen V	gelangen V	15900	bestimmt	bestimmen VVFIN	bestimmt ADJD	415296
doppelt	doppeln V	doppelt A	14837	weiß	weiß ADJD	wissen VVFIN	403674
nahe	nahe PREP	nahen V	14735	führen	fahren|führen VVFIN	führen VVINF	350141
drängen	dringen V	drängen V	14328	lange	lang ADJA	lange ADV	310674
schicken	schick A	schicken V	13169	Bayern	Bayer NN	Bayern NE	301453
nahe	nahen V	nahe A	12338	gehört	gehören|hören VVPP	gehören VVFIN	297240
umgehend	umgehen V	umgehend A	12150	u.	um APPR	und KON	277906
verborgen	verbergen V	verborgen V	12132	überzeugt	überzeugen VVPP	überzeugt ADJD	255330
einführen	einfahren V	einführen V	8474	TuS	unknown ADJA	UNKNOWN NN	250033
koste	kosen V	kosten V	8381	wenig	wenig ADJD	wenige PIS	248789
Morgens	morgen N	morgens ADV	8052	Ihr	Ihr|ihr PPOSAT	ihr PPER	246950
...				...			

Tab. 21: Ambige Wortformen mit mehreren Lemmata, die unterschiedlichen POS-Kategorien angehören

TreeTagger vergibt des Weiteren an Wortformen, denen innerhalb der gleichen Wortart unterschiedliche Lemmata zugeordnet werden können, eine Lemmaform, die diese Mehrdeutigkeit mittels Nennung beider Lemmata zum Ausdruck bringt. Betroffen von der Nennung zweier möglicher Grundformen sind 14.879 Lemmata.

Doppellemma	POS	Frequenz
Ihr|ihr	PRON	1040357
Andrea|Andreas	NE	507781
Franke|Franken	NN	467900
zeihen|ziehen	V	396249
fallen|fällen	V	338773
Reis|Reise	NN	248337
Weste|Westen	NN	239402
Hall|Halle	NN	239334
Esse|Essen	NN	210154
treffen|triefen	V	203174
denken|gedenken	V	199702
fahren|führen	V	199031
fallen|gefallen	V	192563
Eck|Ecke	NN	168603
geraten|raten	V	165890
bieten|gebieten	V	150067
Fall|Falle	NN	124097
gestehen|stehen	V	101007
…		

Tab. 22: Doppellemmata für ambige Wortformen mit mehreren Lemmata innerhalb einer POS-Kategorie (*TreeTagger*)

XIP entscheidet sich bei den Wortformen mit mehreren Lemmatisierungsmöglichkeiten innerhalb der gleichen Wortart konsequent für eine Lemmaform: Die Wortform *Westen* wird immer

dem Lemma *Westen* (und nie der *Weste*) zugeordnet, die Wortform *gedacht* dem Lemma *gedenken* (und nicht dem Lemma *denken*) und die Wortform *Falle* immer dem Lemma *Falle* und niemals dem Lemma *Fall*. *Glemm*, *MS* und häufig auch *MPT* verfahren nach derselben Methode und weisen Wortformen mit mehreren Lemmata immer dasselbe Lemma zu. Die eindeutige Zuweisung einer ambigen Wortform zu einem Lemma kann von NLP-Tool zu NLP-Tool abweichen. *MPT* lemmatisiert die Wortform *gedacht* beispielsweise im Gegensatz zu *XIP* ausschließlich mit dem Lemma *denken*. Die syntaktische Analyse des ganzen Satzes, die die Dependenz-Parser *MS* und *XIP* vornehmen, ließe vermuten, dass Wortformen, die auf mehrere Lemmata abbildbar sind, durch die Auswertung ihres Kontextes disambiguiert werden. Doch der Kontext wird für eine mögliche Disambiguierung bei vielen Wortformen nicht verwendet, sondern eine nicht vorhandene Eindeutigkeit der Abbildung durch eine simplifizierte Zuweisung der betroffenen ambigen Wortformen an immer gleiche Lemmaformen suggeriert.

4.2 POS-Tagging

Nicht nur die Lemmatisierung, sondern auch die Granularität und Nomenklatur der Tagsets ist zwischen den NLP-Tools stark abweichend. Im Artikel von Belica et al. (2011) werden das Inter-Tagger-Agreement, die Fehlertypen und die Fehlerverteilung beim POS-Tagging verschiedener linguistischer Annotationswerkzeuge von DeReKo beschrieben. Belica et al. weisen im Vorwort auf die Tatsache hin, „ [...] that unlike the texts themselves, the morphosyntactic annotations of DeReKo do not have the status of observed data, instead they constitute a theory- and implementation-dependent interpretation“ (2011: 451). Die verschiedenartigen den NLP-Tools zugrunde liegenden linguistischen Theorien und Implementierungen implizieren häufig eine abweichende Annotation der sprachlichen Daten. Im Folgenden wird auf die Unterschiede im POS-Tagging nur anhand von zwei Beispielen eingegangen, der Schwerpunkt dieses Artikels liegt auf der Lemmatisierung.

4.2.1 Mapping der Tagsets

Die Voraussetzung für die Vergleichsmöglichkeit der NLP-Tool-spezifischen Lemmalisten ist die Unifizierung der unterschiedlichen Tagsets. Tagsets mit einer feingranulären Untergliederung für die Wortarten werden auf die allgemeinste Klassifikation abgebildet. Unberücksichtigt bleiben die POS-Tags PREFIX, PUNCT, SYMBOL und TRUNC, die jeweils nur bei einem NLP-Tool Verwendung finden. Durch die Vergabe von Konfidenzwerten für die Güte des POS-Tagging an die einzelnen NLP-Tools können bei Abweichungen der POS-Tags mit dem unifizierten Tagset zu einem bestimmten Lemma die POS-Tags des NLP-Tools mit einem hohen Konfidenzwert in die BLL einfließen (vgl. Anhang).

BLL	*glemm*	*MPT*	*MS*	*TreeTagger*	*XIP*
ADJ	ADJ	A	A	ADJA, ADJD	ADJ
ADV	ADV	ADV	ADV	ADV, PAV, PROAV, PWAV, PTKNEG, PTKA, PTKANT	ADV, NEGAT, PTCL
ART	ART	DET	DET	ART	DET

CONJ	CON	CC,CS	CC,CS	KOKOM, KON, KOUI, KOUS, PTKZU	CONJ
NN	SUB	N	N	NN	NOUN
NUM	NUM	NUM	NUM	CARD	NUM
PREP	PRA	PREP	PREP, POSTP	APPO, APPR, APPRART, APZR	PREP, POSTP
PRON	PRO	PRON	PRON	PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PWAT, PWS	PRON
V	VRB	V	V	VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP, VVFIN, VVIMP, VVINF, VVIZU, VVPP	VERB
ABBR	FUN	N+Abbr	–	–	ABBR
ITJ	–	INTERJ	INTERJ	ITJ	ITJ
NE	EIG	N+Prop	–	NE	NOUN + PROPER
VRZ	VRZ	–	VPART	PTKVZ	–
FOR	FOR	–	–	FM	–
XY	–	–	MSC	XY	–

Tab. 23: Unifikation der Tagsets

4.2.2 Artikel, Determinierer, Pronomen

Im Englischen werden der bestimmte und unbestimmte Artikel unter der Wortart *Determiner* subsumiert, die des Weiteren Demonstrativa, Possessiva und Quantifizierer einschließt, falls diese einem Nomen vorausgehen und über dessen Referenz Auskunft geben. Im Gegensatz zu den NLP-Tools, deren Tagsets primär nach linguistischen Konventionen für das Deutsche ausgerichtet sind (*glemm*, *TreeTagger*), verwenden *MPT*, *MS* und *XIP* nicht das POS-Tag ART für den bestimmten oder unbestimmten Artikel, sondern das POS-Tag DET. *MPT* verwendet das POS-Tag DET wie *glemm* und *TreeTagger* das POS-Tag ART und annotiert nur bestimmte und unbestimmte Artikel mit dem POS-Tag DET. Von *MS* und *XIP* werden neben den Artikeln auch die im Deutschen üblicherweise als attribuierende Pronomen bezeichneten Wortformen (*irgendein*, *solche*, *ebendieser*) mit dem POS-Tag DET ausgezeichnet. Die syntaktische Realisierung (attribuierend oder substituierend) eines Pronomens ist bei *MS* und *XIP* für die Vergabe der POS-Tags DET oder PRON entscheidend. Vor der Unifikation des Tagsets wird das POS-Tag DET in den Lemmalisten von *MS* und *XIP* automatisch auf das POS-Tag PRON umgesetzt, außer für die Lemmata *der,die,das* und *ein*(*e*). Die Umwandlung bestimmter POS-Tags ist notwendig, um die verschiedenartigen linguistischen Annahmen, auf denen die automatische Sprachverarbeitung der NLP-Tools basiert, an einen festgelegten Standard anzupassen.

5. Fazit

Die Erstellung der BLL der nhd. Standardsprache basiert auf großen digitalen Textkorpora, die von mehreren NLP-Tools automatisch linguistisch annotiert sind. Die maschinelle Verarbeitung der Sprachressourcen umfasst die Extraktion der Lemmata, der zugehörigen POS-Tags und Frequenzen aus den annotierten Korpora, die Unifikation der Tagsets der verwen-

deten NLP-Tools, das Alignment unterschiedlicher Nennformen eines Lemmas zwischen den NLP-Tool-spezifischen Lemmalisten, die wahlweise Subsumption nicht lexikalisierter Partizipien und Substantivierungen unter ihrer Grundform und die Rückführung orthographischer Varianten auf eine Lemmaform der aktuellen Standardschreibweise.

Anhand der Unterschiede in der Lemmatisierung der Wortformen im Korpus durch mehrere NLP-Tools wurde gezeigt, dass die automatische Erstellung einer BLL, die lexikographischen Ansprüchen entspricht, alleine mit den automatisch generierten linguistischen Informationen nicht möglich ist. In Abschnitt 4 wurde deutlich, dass eine automatisch generierte Lemmaliste eher durch die Lemmatisierungskonzepte des NLP-Tools beeinflusst ist als durch Korpusbesonderheiten. Lemmalisten, die von verschiedenen NLP-Tools mit dem gleichen Korpus erstellt sind, weichen in Bezug auf die Lemmatisierung stärker voneinander ab, als die von einem NLP-Tool mit verschiedenen (gleich großen) Korpora erstellten Listen. Erst der Vergleich mit der *elexiko*-Stichwortliste garantiert die Zugehörigkeit der Lemmata zum deutschen Wortschatz, valide Lemmaformen und die Vollständigkeit der BLL.

Die morphosyntaktische Analyse in der automatischen Sprachverarbeitung impliziert fehlerhafte und zwischen den NLP-Tools divergierende Lemmaformen und POS-Tags, die falsche Zuordnung von Wortformen zu Lemmata, die Nicht-Erkennung von Wortformen, deren Lemmata folglich fehlen oder die sich aus den Wortformen selbst konstituieren, und die Nicht-Auflösung von Lemma- und POS-Ambiguitäten. Das eigentliche Problem bei der automatischen Erstellung korpusbasierter Lemmalisten stellt nicht die Übernahme von Orthographiefehlern aus dem Korpus dar, sondern das Einfließen von Buchstabenkombinationen, die keine deutschen Lexeme sind, die von den NLP-Tools produzierten Lemmatisierungsfehler, nicht analysierte Wortformen und systematische Abweichungen von der Stichwortkonzeption für Wörterbücher. Für eine Verbesserung der Verfahren in Bezug auf die Lemmatisierung und das POS-Tagging sind die Optimierung der Lexika und morphologischen Analysen sowie die Auswertung des Kontextes relevant.

Wünschenswert wäre eine automatische lexikalische Akquisition, die lexikografischen Standards entspricht. Um dieses Ziel zu erreichen, ist zum einen die Einbindung des lexikografischen Wissens, das in Wörterbüchern steckt, in den automatischen Verarbeitungsprozess der NLP-Tools erforderlich. Lexikografische Informationen über Abweichungen von generellen Lemmatisierungsregeln (z.B. Pluraliatantum, lexikalisierte Partizipien und Superlative) können mit Listen in die Analyse eingebunden werden – die lexikografischen Informationen, die in den Listen kodiert sind, können durchaus auch a priori korpusbasiert gewonnen werden. Zum anderen sollten die Informationen, die die NLP-Tools aus einer Analyse des Satzkontextes gewinnen, in die Auflösung von morphosyntaktischen Ambiguitäten zurückfließen, und nicht nur der Erkennung syntaktischer Funktionen dienen. Nicht immer ist die Disambiguierung heterolemmatischer oder mehreren Wortarten zugehöriger Wortformen über die Einbeziehung des Kontextes möglich, doch wäre das in dem spezifischen Zusammenhang relevante Lemma oder das relevante POS-Tag in vielen Fällen zu bestimmen. Die Möglichkeiten und Grenzen der automatischen Erstellung von korpusbasierten, lexikografisch validen Lemmalisten korrelieren mit dem Aufwand, der in die Implementierung der NLP-Tools investiert wird. Die BLL kombiniert die korpusbasiert ermittelten linguistischen Daten mit lexikografischem Expertenwissen und bietet eine unter lexikografischen Gesichtspunkten optimierte Abbildung der schriftlichen nhd. Standardsprache.

6. Literatur

Die Internetadressen für Forschungsliteratur, Sprachressourcen, Dokumentationen und Projekte beziehen sich auf den Stand von Januar 2014.

Archer, Dawn (Hg.) (2009): What's in a Word-list? Investigating Word Frequency and Keyword Extraction. Farnham.

Baayen, Harald R. (2001): Word Frequency Distributions. Dordrecht.

Belica, Cyril (1994). A German Lemmatizer. Final Report MLAP93-21/WP2. Luxemburg. http://www.ids-mannheim.de/kl/dokumente/glemmrep.pdf.

Belica, Cyril/Kupietz, Marc/Witt, Andreas/Lüngen, Harald (2011): The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/Waßner, Ulrich Hermann (Hg.): Grammatik und Korpora. Dritte Internationale Konferenz. Mannheim, 22.-24.9.2009. Tübingen, S. 451-469. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language 1).

Connexor Oy (1997-2006a): Machinese Language Model. (Internes Dokument).

Connexor Oy (1997-2006b): Machinese Server. (Internes Dokument).

Erb, Sabine (2005): Stichwortansetzung von Partizipien. In: Haß (Hg.), S. 91-95.

Gallmann, Peter (1991): Wort, Lexem und Lemma. In: Augst, Gerhard/Schaeder, Burkhard (Hg.): Rechtschreibwörterbücher in der Diskussion. Geschichte – Analyse – Perspektiven. Frankfurt a. M., S. 261-280.

Guiter, Henri/Arapov, Michail V. (1982): Studies on Zipf's Law. Bochum.

Haß, Ulrike (Hg.) (2005): Grundfragen der elektronischen Lexikographie. Berlin.

Institut für Deutsche Sprache (2011): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2011-I (Release vom 29.03.2011). Mannheim. http://www1.ids-mannheim.de/kl/projekte/korpora/releases.html.

Institut für Deutsche Sprache (2012): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2012-I (Release vom 29.02.2012). Mannheim. http://www1.ids-mannheim.de/kl/projekte/korpora/releases.html.

Keller, Daniela/Schultz, Jörg (2013): Connectivity, Not Frequency, Determines the Fate of a Morpheme. PLoS ONE 8(7). http://www.plosone.org/article/info:doi/10.1371/journal.pone.0069945.

Klosa, Annette (2005): Orthografie und morphologische Varianten. In: Haß (Hg.), S. 133-140.

Klosa, Annette (2011): Von Abbildung bis Wortelement. In: Klosa (Hg.), S. 157-172.

Klosa, Annette (Hg.) (2011): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen. (= Studien zur Deutschen Sprache 55).

Klosa, Annette/Kupietz, Marc/Lüngen, Harald (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica 28, S. 71-98.

Klosa, Annette/Schoolaert, Sabine (2011): Die lexikografische Behandlung von Eigennamen in *elexiko*. In: Klosa (Hg.), S. 193-211.

Knowles, Gerry/Mohd Don, Zuraidah (2004): The notion of a "lemma": Headwords, roots and lexical sets. In: International Journal of Corpus Linguistics 9, 1, S. 69-82. http://www.corpus4u.org/forum/upload/forum/2005071007011220.pdf.

Kühn, Peter (1979): Der Grundwortschatz. Bestimmung und Systematisierung. Tübingen.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the 7th conference on International Language Resources and Evaluation (LREC), Valletta. Valletta (Malta), S. 1848-1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Kupietz, Marc/Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawaguchi, Yuji (Hg.): Working Papers in Corpus-based Linguistics and Language Education, 3. Tokyo, S. 53-59. http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf.

Leech, Geoffry/Rayson, Paul/Wilson, Andrew (2001): Word Frequencies in Written and Spoken English – based on the British National Corpus. Harlow/London.

Müller-Spitzer, Carolin (2005): Die Modellierung lexikografischer Daten und ihre Rolle im lexikografischen Prozess. In: Haß (Hg.), S. 36-54.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. Paderborn.

Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2011): DeReWo: Korpusbasierte Grundformenliste. BLL 0.1 (Release vom 31.12.2011). Mannheim. http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-250000g-2011-12-31-0.1.zip.

Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2012): DEREWO: Korpusbasierte Grundformenliste. BLL 1.0 (Release vom 31.12.2012). Mannheim. http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip.

Pettersson, Helena/Pankow, Christiane (2006): Auswertung der Leistung von zwei frei zugänglichen POS-Taggern für die Annotation von Korpora des gesprochenen Deutsch. Göteborger Arbeitspapiere zur Sprachwissenschaft. https://gupea.ub.gu.se/handle/2077/19368.

Pfeffer, J. Alan (1970): Grunddeutsch. Basic (Spoken) German Dictionary. For Everyday Usage. Englewood Cliffs, NJ.

Popescu, Ioan-Iovitz (2009): Word Frequency Studies. Berlin.

Randall, L. Jones/Tschirner, Erwin (2006): A Frequency Dictionary of German. Core vocabulary for learners. London.

Rosengren, Inger (1969): Wort und Worform. In: Studia Linguistica – Journal of General Linguistics 23, S. 103-113.

Rosengren, Inger (1972-1977): Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt. Süddeutsche Zeitung. 2 Bde. Lund.

Ruoff, Arno (1981): Häufigkeitswörterbuch gesprochener Sprache gesondert nach Wortarten alphabetisch, rückläufig alphabetisch und nach Häufigkeit geordnet. Tübingen.

Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart/Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester. Manchester, UK. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT Workshop, Dublin. Dublin, Ireland. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf.

Schnörch, Ulrich (2002): Der zentrale Wortschatz des Deutschen. Strategien zu seiner Ermittlung, Analyse und lexikografischen Aufarbeitung. Tübingen. (= Studien zur Deutschen Sprache 26).

Schnörch, Ulrich (2005): Die *elexiko*-Stichwortliste. In: Haß (Hg.), S. 71-90.

Seipel, Dietmar/Wegstein, Werner (2011): metaDictionary – Towards a Generic e-Infrastructure for Detecting Variance in Language by Exploiting Dictionary Information. In: Proceedings of the International Symposium on Grids and Clouds, Taipei, PoS(ISGC2011)003. Triest. http://www1.pub.informatik.uni-wuerzburg.de/databases/papers/isgc_2011.pdf.

Seipel, Dietmar/Borek, Luise (2012): A Tool for Cooperative Rule-Based Morpheme Annotation. In: Proceedings of the International Symposium on Grids and Clouds, Taipei, PoS(ISGC2012)012. Triest. http://pos.sissa.it/archive/conferences/153/017/ISGC 2012_017.pdf.

Storjohann, Petra (2005): Das *elexiko*-Korpus. In: Haß (Hg.), S. 55-70.

Tschirner, Erwin (2005): Korpora, Häufigkeitslisten, Wortschatzerwerb. In: Heine, Antje/Hennig, Mathilde/Tschirner, Erwin (Hg.): Deutsch als Fremdsprache – Konturen und Perspektiven eines Fachs. München, S. 133-149. http://www.uni-leipzig.de/herder/red_tools/dl_document.php?PHPSESSID=h6727j48v2a99g3g&id=87&PHPSESSID=h6727j48v2a99g3g.

Wiegand, Herbert Ernst (1983): Was ist eigentlich ein Lemma? Ein Beitrag zur Theorie der lexikographischen Sprachbeschreibung. In: Studien zur neuhochdeutschen Lexikographie, 3 (= Germanistische Linguistik, 82, 1-4), S. 401-474.

Zinsmeister, Heike/Witt, Andreas/Kübler, Sandra/Hinrichs, Erhard (2008): Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In: Lüdeling, Anke/Kytö, Merja (Hg.) Corpus Linguistics. An International Handbook. Bd. 1. Berlin, S. 759-776.

Zipf, George Kingsley (1932): Selected studies of the principle of relative frequency in language. Cambridge, MA.

Zipf, George Kingsley (1935): The Psycho-Biology of Language. An Introduction to Dynamic Philology. Cambridge, MA.

Sprachressourcen, Projekte und Dokumentationen

COSMAS-II:
http://www.ids-mannheim.de/cosmas2/

DEREKO – Deutsches Referenzkorpus:
http://www.ids-mannheim.de/kl/projekte/korpora/

DEREWO – Korpusbasierte Grund-/Wortformenlisten:
http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html

elexiko:
http://www.owid.de/wb/elexiko/start.html

Machinese Phrase Tagger Demo:
http://www.connexor.com/nlplib/?q=demo/mpt

Machinese Syntax Demo:
http://www.connexor.com/nlplib/?q=demo/syntax

Machinese Tagset Morphologie:
http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/morph.html

Machinese Tagset Syntax:
http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/syntax.html

Metalemmaliste:
http://www.metalemmaliste.de/

STTS – Die Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen-Tagset:
http://www.sfs.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html

TIGER-Korpus:
http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

TreeTagger Download:
http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html

Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen – Abbildung von Varianz in Sprache und Genomen:
http://www.sprache-und-genome.de/

Wortschatz-Portal – monolinguale Wörterbücher:
http://corpora.informatik.uni-leipzig.de/

Wortschatz-Portal – Häufigkeitsklassen:
http://asvdoku.informatik.uni-leipzig.de/corpora/index.php?id=fragen-zu-den-suchergebnissen

XIP Dokumentation:
http://open.xerox.com/Services/XIPParser/Pages/Using XIP

XIP Demoseite:
http://open.xerox.com/Services/XIPParser/Consume/64

7. Abkürzungen

BLL	Basislemmaliste
COSMAS	Corpus Search, Management and Analysis System
DEREKO	Deutsches Referenzkorpus
DEREWO	Deutsche Referenzwortlisten
HK	Häufigkeitsklasse
IDS	Institut für Deutsche Sprache
MPT	Machinese Phrase Tagger
MS	Machinese Syntax
nhd.	neuhochdeutsch
NLP	Natural Language Processing
POS	Part-of-Speech
TEI	Text Encoding Initiative
WW-Stichwortliste	Wechselwirkungs-Stichwortliste
XIP	Xerox Incremental Parser
XML	Extensible Markup Language

Anhang: Die Verarbeitung der Sprachressourcen – Programmmodule

Die Erstellung der BLL der nhd. Standardsprache ist auf mehrere aufeinanderfolgende Schritte verteilt, in denen die Sprachressourcen (vgl. Abschnitt 3) automatisch verarbeitet werden. Die Module 1 bis 4 verarbeiten die annotierten Korpora und die daraus extrahierten Lemmalisten separat.

1. Extraktion der relevanten Daten (Lemma, POS-Tag, Frequenz) aus den linguistisch annotierten Korpora, Erstellung der NLP-Tool-spezifischen „Ur"-Lemmalisten
2. Vereinheitlichung, Bereinigung und Korrektur der Lemmata, Unifikation der Tagsets
3. Abgleich und Alignierung der Lemmalisten mit der WW-Stichwortliste, Erzeugung der NLP-Tool-spezifischen Basis- und Restlemmalisten
4. Erzeugung der NLP-Tool-spezifischen Eigennamenlisten

Modul 5 übernimmt die Zusammenführung der NLP-Tool-spezifischen BLLen und Eigennamenlisten zu (einer) finalen Version(en).

5. Merging der NLP-Tool-spezifischen BLLen und Eigennamenlisten

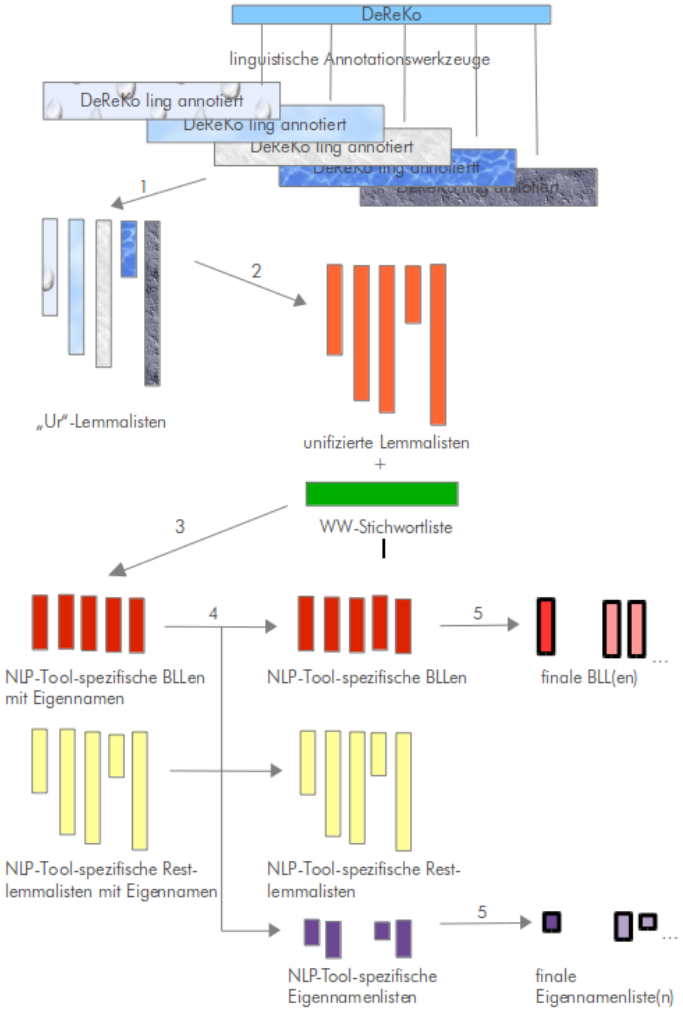

Abb. 5: Die Programmmodule

Durch den Vergleich der NLP-Tool-spezifischen Lemmalisten untereinander und mit der WW-Stichwortliste wird automatisch eine Lemmaliste mit POS-Tags und Frequenzangaben generiert, die mit lexikografischen Ansprüchen korreliert. Die händisch erbrachte Arbeit konnte sich überwiegend auf die Verifizierung frequenter Lemmata aus dem Korpus (ca.

25.000), die nicht in *elexiko* enthalten sind, und die Programmierung konzentrieren. Die Verarbeitung wurde mit Perl- und Shell-Skripten realisiert, die Ergebnisse der einzelnen Verarbeitungsschritte werden automatisch dokumentiert und die Lemmalisten nach den einzelnen Schritten zwischengespeichert. Anhand dieser Ergebnisse zeigen sich Spezifika der einzelnen NLP-Tools, die in Abschnitt 4 exemplarisch ausgeführt sind.

Die Programme sind modular aufgebaut, das Korpus und die linguistischen Annotationswerkzeuge können ausgetauscht oder neue Sprachressourcen integriert werden. Der Vergleich mit der WW-Stichwortliste kann auch übersprungen und eine BLL allein durch die Komparation der aus den annotierten Korpora extrahierten NLP-Tool-spezifischen Lemmalisten generiert werden. In diesem Fall kann neben der Frequenz nur das Vorkommen einer Lemma-POS-Tag Kombination in mehreren NLP-Tool-spezifischen Lemmalisten als Kriterium für seine Richtigkeit und damit für die Aufnahme in die BLL dienen. Bestimmte regelhafte Lemmatisierungsfehler und die Verwendung von Wortformen aus Flexionsreihen als zusätzliche Lemmata werden jedoch von sämtlichen hier untersuchten NLP-Tools praktiziert (vgl. Abschnitt 4.1), wodurch Sprachmaterial, das nicht der nhd. Standardsprache angehört, in der BLL stehen würde.

Die Korpusextraktion der Lemmalisten (Modul 1) verläuft NLP-Tool-spezifisch aufgrund der stark abweichenden Granularität der linguistischen Annotationen und der unterschiedlichen XML- bzw. XCES-Auszeichnung der einzeln annotierten Korpora. Die Frequenzen der Lemmata mit dem dazugehörigen POS-Tag werden addiert und in NLP-Tool-spezifischen frequenzsortierten Lemmalisten subsumiert. Ein Vergleich der häufigsten Lemmata in den frequenzsortierten NLP-Tool-spezifischen „Ur“-Lemmalisten verdeutlicht die Notwendigkeit der Vereinheitlichung von POS-Tags und bestimmten Lemmata (Artikel und Pronomen):

glemm			*MPT*			*TreeTagger*		
d-	ART	372041251	die	DET	195144842	die	ART	348613119
in	PRA	102571346	der	DET	140798904	,	$,	220936151
und	CON	84379423	in	PREP	100152883	.	$.	209869299
sein	VRB	59777467	.	V	92956679	und	KON	85672890
werden	VRB	41845240	und	CC	85465442	UNKNOWN	NE	78522261
von	PRA	40739636	,	N	77073025	eine	ART	75238859
mit	ADV	35192255	,	V	69769840	in	APPR	66855760
haben	VRB	34950734	.	N	65259304	"	$(	64893662
ein	PRO	33762990	sein	V	64537359	unknown	NN	61548996
an	PRA	31994097	das	DET	63511136	sein	VAFIN	59478241
zu	FUN	31115731	zu	PREP	45016087	@card@	CARD	34578550
für	PRA	26860147	ein	DET	42900667	von	APPR	34547607
er/sie/es/sie	PRO	26382736	werden	V	41814302	mit	APPR	33822577
auf	PRA	25893375	von	PREP	38781798	haben	VAFIN	33506610
nicht	ADV	23019130	die	PRON	36974814	im	APPRART	30947199
ein	ART	21870507	haben	V	35623879	:	$.	29864399
eine	NUM	21323830	,	NE	29851099	werden	VAFIN	29848230
es	PRO	21016363	eine	DET	29293374	für	APPR	27937099
auch	ADV	20274147	mit	PREP	28251627	er|es|sie	PRF	26745646
bei	PRA	19374426	für	PREP	28099826	auf	APPR	24950777
zu	PRA	17978658	sich	PRON	26734882	)	$(	24493758
sie	PRO	17907983	an	PREP	26043773	(	$(	24455221
er	PRO	17644340	auf	PREP	24814732	nicht	PTKNEG	23471015
als	PRA	17589707	es	PRON	23932297	die	PRELS	23210571
sein	PRO	15404538	nicht	ADV	23410488	es	PPER	21552762
nach	PRA	15182829	"	N	21870500	auch	ADV	20529329
…			…			…		

Tab. 24: Frequenzsortierte NLP-Tool-spezifische „Ur“-Lemmalisten

Von Modul 2 werden mehrere Schritte zur Vereinheitlichung und Korrektur der Lemmalisten ausgeführt:

- Extraktion von Lemmata, die Sonderzeichen (@,#,www) oder keine alphabetische Zeichen (Zahlen, Satzzeichen) enthalten
- Überführung der Lemmata von *MPT* und *XIP* durch regelbasierte Algorithmen auf eine in der Lexikografie übliche Form[31]
- Vereinheitlichung der Lemmaformen für Pronomen und Artikeln mittels einer händisch erstellten Liste
- Unifikation der Tagsets
- Ermittlung des Konfidenzwertes für die Güte der POS-Tags (vgl. Modul 5)
- Korrektur sehr häufiger falscher Lemmata und POS-Tags mittels einer händisch erstellten Liste
- Generierung von Testlisten: ambige Lemmata (ursprüngliches und unifiziertes Tagset), Tagset mit Häufigkeit der POS-Tags (ursprüngliches und unifiziertes Tagset)

glemm			*MPT*			*TreeTagger*		
der,die,das	DET	372041251	der,die,das	DET	399454882	der,die,das	DET	350956922
in	PREP	102571346	in	PREP	100152883	und	CONJ	85672890
und	CONJ	84379423	und	CONJ	85465442	ein(e)	DET	75834874
sein	V	59777467	ein(e)	DET	72201801	in	PREP	66870303
werden	V	41846394	der,die,das	PRON	67395083	sein	V	64499919
von	PREP	40739636	sein	V	64537359	werden	V	41819359
mit	PREP	35192255	zu	PREP	45016087	haben	V	35386600
haben	V	34950734	werden	V	41814302	der,die,das	PRON	34973830
ein(e)	PRON	33762990	von	PREP	38781798	von	PREP	34553397
an	PREP	31994097	haben	V	35623879	mit	PREP	33822577
zu	ADV	31115731	mit	PREP	34859900	im	PREP	30947199
für	PREP	26881002	für	PREP	28099826	für	PREP	27949705
sich	PRON	26341487	sich	PRON	26734882	sich	PRON	26745646
auf	PREP	25893375	an	PREP	26043773	auf	PREP	24955025
ein(e)	DET	25776090	auf	PREP	24814732	nicht	ADV	23471015
nicht	ADV	23019130	es	PRON	23932297	es	PRON	21552762
ein(e)	NUM	21329931	nicht	ADV	23410488	auch	ADV	20529329
es	PRON	21016363	auch	ADV	19599208	er	PRON	20437336
auch	ADV	20274147	sie	PRON	19576575	zu	CONJ	19307431
bei	PREP	19374426	er	PRON	18331833	als	CONJ	17354281
sein	PRON	18596588	als	CONJ	17870326	am	PREP	16497099
zu	PREP	17978658	bei	PREP	17368189	sein	PRON	15761347
sie	PRON	17907983	sein	PRON	15559970	dies(e,er,es)	PRON	15329952
er	PRON	17644340	dies(e,er,es)	PRON	15421438	bei	PREP	14377464
als	CONJ	17589707	dass	CONJ	14242391	nach	PREP	14136194
nach	PREP	15182829	ihr	PRON	14118537	können	V	14045367
dies(e,er,es)	PRON	14816130	können	V	14103454	sie	PRON	13647883
aus	PREP	14772649	Jahr	NN	11794241	ihr	PRON	13577927
…			…			…		

Tab. 25: Frequenzsortierte NLP-Tool-spezifische Lemmalisten nach Durchlaufen des Moduls 2

[31] Die internen Morphologiekomponenten von *MPT* und *XIP* führen einzelne Morpheme in Komposita auf ihre (vermeintlichen) Grundformen zurück und bilden aus ihnen das Lemma (*haupt frieden hof* 'Hauptfriedhof', *alter grund* 'Altersgründe', *schule abschluss – schulen#abschluss* 'Schulabschluss', *dunkeln#blau* 'dunkelblau', *mit=gliedern#Versammelung* 'Mitgliederversammlung', *fachen#Hochschule* 'Fachhochschule', *an=liegen}heit* 'Angelegenheit'). *MS* stellt die Bestandteile, aus denen Komposita bestehen, in der Lemmaform lediglich durch eine Raute getrennt und den Anfangsbuchstaben als Minuskel dar (*haupt#fried#hof, dunkel#blau*).

Modul 3 übernimmt den Abgleich der NLP-Tool-spezifischen Lemmalisten mit der WW-Stichwortliste. Die BLL soll keine Lemmata enthalten, die es im Deutschen nicht gibt. Lemmatisierungsfehler, Orthographiefehler oder fremdsprachliche Belegstellen führen ebenso zu Lemmata in den Lemmalisten, die in keinem Wörterbuch der nhd. Standardsprache stehen, wie die vom lexikografischen Standard abweichenden Lemmatisierungskonzepte der NLP-Tools (vgl. Abschnitt 4.1). Die in der WW-Stichwortliste nicht enthaltenen Partizipien und Substantivierungen werden automatisch mit den in der WW-Stichwortliste verzeichneten Grundformen aligniert und die Frequenzen zu denen der Grundform addiert. Übersteigt die Frequenz von Substantivierungen (*Aufständische*, *Verantwortliche*, *Hinterbliebene*, *Nachsehen*) oder adjektivisch gebrauchten Partizipien (*abschließend*, *nachfolgend*, *aufregend*, *rauschend*, *eingeschlossen*) die Frequenz der Grundform, werden die abgeleiteten Formen zusätzlich in die BLL aufgenommen, der sprachliche Usus soll sich in der BLL widerspiegeln (vgl. Abschnitt 4.1.2). Von Modul 3 werden folgende Schritte beim Abgleich der NLP-Tool-spezifischen Lemmalisten mit der WW-Stichwortliste ausgeführt:

– exakter Stringvergleich
– Mapping der orthographischen Varianten auf die Standardschreibweise
– Alignierung ss-ß
– Alignierung Paritzip1
– Alignierung Partizip2
– Alignierung Substantivierungen
– Alignierung Pluraliatantum

Durch den Abgleich mit der WW-Stichwortliste entstehen NLP-Tool-spezifische BLLen, die lexikografisch validierte Lemmata enthalten. Daneben werden NLP-Tool-spezifische Rest lemmalisten gebildet, in die diejenigen Lemmata einfließen, die nicht in der WW-Stichwortliste verzeichnet sind und die nicht aligniert werden können. Die BLL wächst in einem iterativen Prozess: Durch die händische Bearbeitung der frequenzsortierten Restlemmalisten werden neue Lemmata verifiziert und in die WW-Stichwortliste integriert oder neue Alignierungsalgorithmen implementiert. Die betroffenen Lemmata stehen nach dem nächsten Programmlauf nicht mehr in der Restlemmaliste, sondern in der BLL.

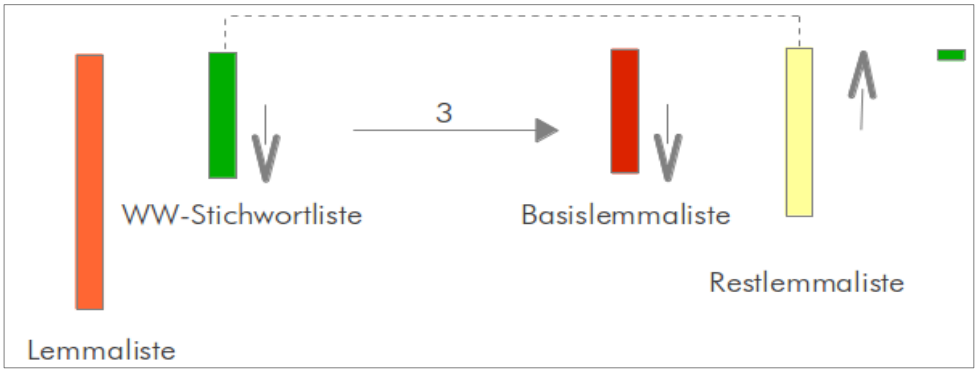

Abb. 6: Programmmodul 3

In der Restlemmaliste befinden sich sowohl Lemmata, die nach einer händischen Überprüfung oder der Implementierung weiterer Alignierungsalgorithmen in die BLL einfließen, als auch sprachliches Material, das unter lexikografischen Gesichtspunkten nicht in die BLL übernommen wird. Hauptsächlich im vorderen Bereich der frequenzsortierten Restlemmalisten stehen viele Lemmata, die spezifische Lemmaformen der einzelnen NLP-Tools

darstellen (*Fall|Falle*, *BERLIN*, *Usa*, *sondern auch*), falsche Lemmatisierungen (*hamburger*, *Salzbürger*, *unbedingen*, *Landspiel*, *unverändern*, *gottloben*), Orthographiefehler (*wiedersprechen*, *Prophezeihung*, *schießlich*) und Wortformen, die über die Alignierung noch nicht auf ihre Grundform zurückgeführt wurden (*Große*, *Ganze*, *Gleichzeitig*, *Seinerzeit*).

Die verbleibenden lexikografisch interessanten Lemmata in den Restlemmalisten verteilen sich überwiegend auf Komposita, Neologismen und Fremdwörter (*Abfallsünder*, *Airshow*, *alternativlos*, *Aquafitness*, *Ärztemangel*, *Befristungsgesetz*, *Bergleute*, *berufsübergreifend*, *Bitfolge*, *Boyband*, *Brustwirbelsäule*, *deinstallieren*, *Ferienplanung*, *googeln*, *Handelsstreitigkeit*, *Hedgefonds*, *Hochseilgarten*, *Internetblase*, *jammen*, *Kleingewerbetreibende*, *Königsposition*, *Kopftuchträgerin*, *Kostenfalle*, *Mentorenprogramm*, *Mixtape*, *Onlinebroker*, *Opferdiskurs*, *Orchesterhaus*, *Priesterstand*, *Radiomultikulti*, *Ratingsystem*, *Ravekultur*, *Rolltor*, *Romafamilie*, *Schiedsrichterkosten*, *Seniorenrunde*, *spacig*, *Stabmixer*, *Trefferwahrscheinlichkeit*, *Trinkwasserprojekt*, *Voting*, *Weblink*, *Wildschweinbestand*, *Wohnungslose*), daneben sind Herkunftsbezeichnungen (*Lampertheimer*, *Ludwigshafener*, *Yorker*, *tschetschenisch*), Eigennamen (*Carlos*, *Black*, *Romanshorn*, *Strobl*, *Hillary*) und Movierungsvarianten (*Neugeborene*, *Geschworene*, *Kriminalbeamte*) vertreten. Durch die sukzessive händische Validierung der lexikografisch relevanten häufigen Lemmata aus den Restlemmalisten wird die BLL aktualisiert und an die Lexik der sich aus der Korpuszusammensetzung ergebenden Themen angepasst.

Modul 4 extrahiert die Eigennamen aus den NLP-Tool-spezifischen BLLen und Restlemmalisten und fasst sie in NLP-Tool-spezifischen Eigennamenlisten zusammen. Ein zusätzliches POS-Tag für Eigennamen vergeben *glemm*, *MPT*, *TreeTagger* und *XIP*. Die Anzahl der Eigennamen und die Eigennamen selbst sind zwischen den NLP-Tools stark different. Für die Aufnahme in die Eigennamenliste 1.0 muss ein Eigenname von mindestens zwei NLP-Tools als solcher erkannt werden, eines der beiden NLP-Tools, die den Eigennamen verzeichnen, muss *glemm* (28.817 Eigennamen) oder *TreeTagger* (36.769 Eigennamen) sein. Dadurch wird die Anzahl von 4.402.071 Eigennamen bei *MPT* bzw. 6.277.930 Eigennamen bei *XIP* erheblich reduziert. *MPT* und *XIP* annotieren nicht analysierbare nominale Tokens (*Reissnägel*, *Baltzerdie*, *Ballstopper*, *Assisstent*) und längere Komposita (*Affekt-Kontroll-Training*, *Ballettvorbereitung*) sehr häufig als Eigennamen, obwohl es sich um Tokenisierungs-, Orthographiefehler oder Gattungsnamen handelt. Über zusätzliche manuell erstellte Listen, die nur Nomen enthalten, die ausschließlich als Eigennamen fungieren, werden nicht erkannte (und folglich mit dem POS-TAG NN ausgezeichneten) Eigennamen aus den Lemmalisten extrahiert und das POS-Tag bei der Integration in die Eigennamenliste zu NE verändert. Mit Mustererkennung (*straße*, *gasse*, *bach*, *hausen*) werden weitere Eigennamen in den Lemmalisten gefunden, die das POS-Tag NN führen. Die Mustererkennung wird nur auf die Restlemmaliste angewandt, in der BLL sind überwiegend Gattungsnamen mit den Mustern enthalten (*Umgehungsstraße*, *Schnellstraße*, *Wasserstraße*). Die nach diesen Kriterien generierte Eigennamenliste 1.0 verzeichnet 68.892 Einträge, eine Eigennamenliste, die ohne Mustererkennung erstellt wird, verzeichnet 39.261 Einträge.

Zahlreiche Nomen stehen sowohl in der BLL mit dem POS-Tag NN als auch in der Eigennamenliste mit dem POS-Tag NE. Eine ambige Verwendungsweise von Nomen im Deutschen tritt durch die Gleichlautung von Gattungsbezeichnungen und Eigennamen häufig auf (*Müller NN 352771*, *Müller NE 539549*). Das Format für die Eigennamenliste entspricht dem Format der BLL (vgl. Abschnitt 2.3), die Häufigkeitsklassen werden mit der Frequenz des häufigsten Lemmas der BLL berechnet (vgl. Abschnitt 2.4).

Rang	Lemma	POS	Häufigkeitsklasse
1	Deutschland	NE	7
2	Berlin	NE	8
3	Peter	NE	8
4	Michael	NE	9
5	Thomas	NE	9
6	Österreich	NE	9
7	München	NE	9
8	Europa	NE	9
9	Bad	NE	9
10	Wien	NE	9
11	Hans	NE	9
12	Wolfgang	NE	9
13	Frankfurt	NE	9
14	Martin	NE	9
…			
68882	Zwergengasse	NE	29
68883	Zwergenstraße	NE	29
68884	Zwerghausen	NE	29
68885	Zwiebackshausen	NE	29
68886	Zwiefaltbach	NE	29
68887	Zwingenbach	NE	29
68888	Zwinggasse	NE	29
68889	Zwirnereigasse	NE	29
68890	Zwischenposthofstraße	NE	29
68891	Zwölfschlösserhausen	NE	29
68892	Zwoschwitzbach	NE	29

Tab. 26: Eigennamenliste 1.0

Der Abgleich der NLP-Tool-spezifischen BLLen oder Eigennamenliste in Modul 5 garantiert ein Maximum an Informationen zu einem Lemma, dessen POS-Tag(s) und Häufigkeit im Korpus. Die NLP-Tool-spezifischen Lemmalisten unterscheiden sich neben den Lemmata auch in den POS-Tags zu den einzelnen Lemmata und den akkumulierten Frequenzen. Es gibt keine NLP-Tool-spezifische Lemmaliste, die für Lemmata und POS-Tags hundertprozentig verlässliche Ergebnisse (Precision) und eine komplette Abdeckung der Wortformen im Korpus (Recall) bietet. Eine BLL der nhd. Standardsprache variiert nicht nur mit der Korpusgrundlage, sondern vor allem mit den ausgewählten NLP-Tools, die bei der linguistischen Annotation des Korpus Verwendung finden, sowie dem Aufwand der lexikografischen Nachbearbeitung.

Das Aufnahmekriterium für ein Lemma in die BLL ist sein Vorhandensein in der WW-Stichwortliste. Ausschlaggebend für die Übernahme eines POS-Tags für ein bestimmtes Lemma in die BLL ist das Vorkommen des POS-Tags mit dem betreffenden Lemma in mehreren NLP-Tool-spezifischen BLLen. Weichen die POS-Tags für ein Lemma in den NLP-Tool-spezifischen Lemmalisten voneinander ab, werden die POS-Tags des NLP-Tools mit einem hohen Konfidenzwert für das POS-Tagging übernommen. Der Konfidenzwert für das POS-Tagging der einzelnen NLP-Tools wird anhand des manuell ausgezeichneten TIGER-Korpus ermittelt.[32] In die Berechnung des Konfidenzwertes fließt ein, wie viele der POS-Tags der NLP-Tool-spezifischen Lemmalisten jeweils mit den POS-Tags für die Lemmata aus dem TIGER-Korpus identisch sind und wie viele POS-Tags davon abweichen.

[32] TIGER-Korpus: www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html.

Während des Vergleichs der NLP-Tool-spezifischen BLLen wird ein Mittelwert aus den teilweise stark abweichenden absoluten Frequenzen für ein Lemma mit dem zugehörigen POS-Tag berechnet, der in der finalen BLL verzeichnet ist. Als finale BLL wird diejenige Version gewählt, die aus dem Vergleich der vier BLLen von *Glemm*, *MPT*, *MS* und *TreeTagger* entsteht (die Annotationen von XIP waren zum Entstehungszeitpunkt der BLL 1.0 nicht verfügbar). Auch die Zusammenführung von nur zwei oder drei NLP-Tool-spezifischen BLLen oder Eigennamenlisten ist in Modul 5 möglich und wird zu Vergleichszwecken durchgeführt.